Research Article

Statistics in Medicine

Received 3 June 2011, Accepted 14 October 2011 Published online 11 January 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4462

A two-stage Bayesian design for co-development of new drugs and companion diagnostics

Stella Wanjugu Karuri and Richard Simon*†

Most new drug development in oncology is based on targeting specific molecules. Genomic profiles and deregulated drug targets vary from patient to patient making new treatments likely to benefit only a subset of patients traditionally grouped in the same clinical trials. Predictive biomarkers are being developed to identify patients who are most likely to benefit from a particular treatment; however, their biological basis is not always conclusive. The inclusion of marker-negative patients in a trial is therefore sometimes necessary for a more informative evaluation of the therapy. In this paper, we present a two-stage Bayesian design that includes both marker-positive and marker-negative patients in a clinical trial. We formulate a family of prior distributions that represent the degree of *a priori* confidence in the predictive biomarker. To avoid exposing patients to a treatment to which they may not be expected to benefit, we perform an interim analysis that may stop accrual of marker-negative patients or accrual of all patients. We demonstrate with simulations that the design and priors used control type I errors, give adequate power, and enable the early futility analysis of test-negative patients to be based on prior specification on the strength of evidence in the biomarker. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: clinical trials design; predictive biomarkers; Bayesian inference; prior distribution; type I error probabilities

1. Introduction

A majority of new drug development in oncology is molecular target based. Because of the heterogeneity of tumors of a primary site, treatment benefit is likely for only a subset of patients in clinical trials based on primary site. A typical example is Herceptin (Genentech Inc., San Francisco, CA), which is used for the treatment of breast cancer patients whose tumors overexpress the HER2 gene. Predictive biomarkers are a class of biomarkers that identify patients who are likely to benefit from a particular treatment. Establishing medical utility of a previously developed candidate predictive biomarker generally requires a clinical trial in which treatment assignment is randomized between the treatment and control [1]. One design strategy for using a predictive biomarker is the enrichment design [2], where marker-negative patients are excluded from the study. Excluding marker-negative patients from the trial has been shown to be more efficient than standard designs [3, 4] where the treatment effect in marker-negative patients is substantially less than in marker-positive patients and the prevalence of the marker positives is less than 50%. In many cases, however, the biological basis for believing that the candidate biomarker will properly distinguish those patients who benefit from the new treatment from those who will not is not conclusive [1], and it is desirable to include marker-negative patients in the trial [5, 6].

Wang *et al.* [7] developed a two-stage adaptive design that includes marker-positive and marker-negative patients but provides for termination of recruitment to the marker-negative stratum based on an interim analysis. Freidlin and Simon [8] and Jiang *et al.* [9] have proposed other adaptive designs. Here, we use a Bayesian formulation to provide a flexible framework for representing the degree of prior

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA
**Correspondence to: Richard Simon, Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA.*
†*E-mail: rsimon@mail.nih.gov*

confidence in the biomarker. This degree of confidence may range widely in different clinical development contexts. In some cases, there will be relatively strong confidence in the biomarker but not strong enough to exclude marker-negative patients, or regulators may insist on the inclusion of marker-negative patients to obtain data supporting the approval of the test. In such a trial, however, it will be important to avoid including too many marker-negative patients and exposing them to a treatment to which they are not expected to benefit based on prior evidence. Although Bayesian methods provide flexibility for incorporating prior confidence in the candidate predictive biomarker, a major challenge is developing a Bayesian formulation that also provides robust inferences that meet the standards for conventional frequentist phase III clinical trials.

Section 2 details our formulation of the problem. A vast body of work is available on prior specification with regard to clinical trials [10–13]. We describe a class of two-point priors to describe treatment effect and how it varies based on biomarker value. We utilize these simple two-point priors to facilitate understanding on the operating characteristics of the designs we propose and indicate later how the designs can be extended to a broader set of priors. We used a modified Bayesian version of the two-stage design of Wang *et al.* [7] in Section 3. Unlike Berry's [10] approach where decisions on patient recruitment after the interim stage are made via the predictive distribution, we use the posterior distributions of treatment effect within biomarker strata. Simulation studies in Section 4 enable an empirical study of type I error, power, and sample size. In Section 6, we discuss results, suggest extensions, and provide concluding remarks.

2. Methodology

We take survival as the endpoint and assume the data to follow a proportional hazard model within the marker-positive and marker-negative strata separately

$$\log(h(t)/h_0(t)) = \begin{cases} \delta_+, & \text{for test-positive patients} \\ \delta_-, & \text{for test-negative patients,} \end{cases} \tag{1}$$

where δ_+ and δ_- are the treatment effects in the test-positive and test-negative patients, respectively. Given the data D and the test-positive and test-negative treatment effects δ_+ and δ_-, the observed treatment effects are taken to be independent and approximately Normal:

$$\hat{\delta}_+|\delta_+, D \sim N(\delta_+, 4/E_+), \tag{2}$$

$$\hat{\delta}_-|\delta_-, D \sim N(\delta_-, 4/E_-). \tag{3}$$

The distribution assumption are based on the approximation of the log-rank statistics and its asymptotic distribution. The parameters E_+ and E_- are the number of events in the test-positive and test-negative patients at the time of analysis. Prior information on the mean treatment effects δ_+ and δ_- is specified via a distribution denoted as $P(\delta_+, \delta_-)$.

2.1. *Prior distributions*

Although we use a Bayesian formulation, because we are proposing a phase III trial design that may be used for regulatory or practice standard decision making, the type I error, the error of concluding a treatment work when it does not, must be controlled. Three hypotheses are relevant when using a predictive biomarker:

1. $H_0 : \delta_+ = 0$ denoted as H_{0+},
2. $H_0 : \delta_- = 0$ denoted as H_{0-}, and
3. $H_0 : \delta_+ = \delta_- = 0$ denoted as H_{0+-}.

We specify a four-point prior distribution with mass on the points $\{(\delta_+, \delta_-) = (0,0), (\delta_*, 0), (0, \delta_*), (\delta_*, \delta_*)\}$, where δ_* is the log-hazard ratio for treatment effect which is of clinical importance to detect. The parameter δ_* is derived from prior preclinical, phase I, or phase II study where the treatment has shown promise of efficacy. In frequentist designs, the number of events in the trial are planned to detect an alternative with a log-hazard ratio of δ_* for a specific power and significance level. δ_* is therefore a natural choice for an alternative outcome to the marginal null.

We parameterize the prior as follows:

$$P(\delta_+ = 0|\delta_- = \delta_*) = r_2, \tag{4}$$

$$P(\delta_- = 0|\delta_+ = \delta_*) = r_1, \tag{5}$$

$$P(\delta_+ = 0, \delta_- = 0) = p_{00}. \tag{6}$$

The parameter r_2 represents the prior probability of having no treatment effect in the test positives when a treatment effect does exist in the test negatives. The parameter r_1 represents the prior probability of having no treatment effect in the test negatives when a treatment effect does exist in the positives. This parametrization offers two key advantages. The first being that the overall type I error for H_{0+-} is controlled by p_{00}. The other advantage is that prior information on the classifier can be incorporated through specification of r_1 and r_2. For targeted therapy, confidence in the classifier means belief that a treatment effect exists in test-positive patients only. A large value of r_1 and a small value for r_2 would represent this belief. Conversely, little or no confidence in the classifier would imply that if a treatment effect exists, it exists in both test-positive and test-negative patients; small values of r_1 and r_2 would represent such a belief.

The posterior probabilities related to the formulated hypothesis can be shown as follows (Appendix A):

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_1 L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0} + aq_2 L_{0,\delta_*}}\right)^{-1}, \tag{7}$$

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2 L_{0,\delta_*} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0} + aq_1 L_{\delta_*,0}}\right)^{-1}, \tag{8}$$

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2 L_{0,\delta_*} + aq_1 L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0}}\right)^{-1}, \tag{9}$$

where $a = (1 - r_1)(1 - r_2)/(1 - r_1 r_2)$, $q_1 = r_1/(1 - r_1)$, $q_2 = r_2/(1 - r_2)$, $q_{00} = p_{00}/(1 - p_{00})$, and $L_{i,j} = P(\hat{\delta}_+, \hat{\delta}_-|\delta_+ = i, \delta_- = j)$ denote the marginal conditional probabilities of the bivariate Normal density given in Equation (3). Complete *a priori* confidence in the utility of the classifier means $r_1 \to 1$ and $r_2 \to 0$, and the posterior probability of the treatment failing in the test negatives given in Equation (8) will be large and unaffected by p_{00}. However, the posterior probability of the treatment failing in the test positives in Equation (7) will be small if p_{00} is small and large if p_{00} is large. Complete lack of confidence in the classifier means $r_1 \to 0$ and $r_2 \to 0$; the treatment is equally likely to work in the test positives and test negatives, and the posterior probabilities given in Equations (7) and (8) are approximately equal. Under these conditions, the posterior probability of the treatment having no effect in the test negatives and test positives given in Equation (9) is largely determined by the values of p_{00} and the data.

Values for r_1 and r_2 are set by the investigator based on their belief in the classifier. However, p_{00} is chosen to control type I error. The simulation exercise in Section 4 outlines a method of obtaining p_{00} by examining the empirical probabilities given in Equations (7)–(9).

3. Trial design

We utilize a Bayesian version of the adaptive design of Wang *et al.* [7] by considering a two-stage design where the posterior distribution of treatment effects within the marker-positive and marker-negative strata based on an interim analysis of first-stage data is used to make decisions on patient recruitment and trial progression. The trial is prospectively sized for detecting a treatment effect in test-positive patients using a target treatment effect of δ_*, a power of $(1-\beta)$, and a two-sided α significance level. The required number of test-positive events is

$$E_+ = (4(z_\alpha + z_\beta)^2)/(\delta_*^2).$$

For a nonprognostic classifier, the expected number of test-negative events is approximately

$$E_- = ((1 - \text{prev})/\text{prev})E_+,$$

where prev denotes the prevalence of marker positivity. We size the clinical trial on the basis of the number of observed events in the test-positive stratum at the final analysis. The prevalence of positivity is important as it determines the rate of accrual of marker-negative patients relative to marker-positive patients. Prevalence therefore influences the operating characteristics of the design, particularly the power of detecting a treatment effect in the test-negative stratum. Prevalence of marker positivity is usually available prior to initiation of phase III trials from archival tissues from patients with the target disease.

At the interim analysis, a proportion t of E_+ and E_- events are used to estimate the treatment effects, and the posterior probabilities given in Equations (7)–(9) are calculated. For a particular threshold value TH, the following sequence of actions are taken:

1. Stop the trial if $P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) \geq TH$ at the interim stage;
2. stop negative recruitment at the interim stage if $P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) \geq TH$; or
3. continue recruitment of both test-positive and test-negative patients.

The first course of action is necessitated by the fact that the trial therapy is a targeted therapy. If the treatment is ineffective in test-positive patients, it is unlikely to be effective in the test-negative patients and the trial should be stopped. The next course of action ensures that recruitment is stopped for test-negative patients if the therapy is ineffective in this subpopulation. To reduce the number of parameters in the design, we use the same TH value. However, different values of TH can be used in steps 1 and 2. Ideally, the threshold should be a large value. However, a very large value will reduce the effectiveness of futility monitoring at the interim stage, which may compromise patient safety. For example, if the treatment is indeed ineffective in the marker-positive and marker-negative patients, having a threshold value close to one reduces the probability of stopping the trial, which results in the continued accrual of patients, exposing patients to toxic side effects of an ineffective treatment. We use a sequence of TH values to assess the performance characteristic of the design. At the final analysis, the hypothesis H_{0+} is rejected if $P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) < \epsilon$, where ϵ is some small value. Similarly, H_{0-} is rejected if $P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) < \epsilon$. We use $\epsilon = 0.05$.

4. Simulation study

Using prevalence values of 25% and 50% for marker-positive patients, we simulated data under the following scenarios:

1. No treatment effect, with $(\delta_+, \delta_-) = (0, 0)$;
2. An observed treatment effect in test-positive patients only with $(\delta_+, \delta_-) = (\delta_*, 0)$;
3. The unlikely scenario where there is a treatment effect in the test-negative patients but none in the test-positive patients, with $(\delta_+, \delta_-) = (0, \delta_*)$; and
4. An observed treatment effect in the test-positive and test-negative patients with $(\delta_+, \delta_-) = (\delta_*, \delta_*)$,

We used values of $\delta_* = \log(1/2)$ and $\delta_* = \log(2/3)$, which correspond to a one-half and one-third reduction in hazard attributed to the treatment. Most phase III studies in oncology are designed to detect a reduction in hazard of approximately between 25% and 33% [14–17]. A 50% reduction in hazard is considered large but has been observed in a predictive biomarker study [18]. For a stratified design with $\delta_* = \log(2/3)$, the number of events required for the test positives is $E_+ = 256$. For prevalence of 50% and 25%, this implies $E_- = 256$ and $E_- = 768$, respectively. If $\delta_* = \log(1/2)$, the number of events for test positives is $E_+ = 88$ with resulting $E_- = 88$ and $E_- = 164$ for a 50% and 25% prevalence, respectively.

The asymptotic distribution of the log-rank statistics at the interim and final stages is bivariate Normal with unit variances and correlation equal to $\rho = \sqrt{t} = \sqrt{E_1/E_2}$, where E_1 and E_2 are the numbers of events in the interim and final stage, respectively [19]. Consequently, simulations for the observed test-positive effects at the interim and final stage can be obtained using a bivariate Normal distribution, where the interim and final effects for the test positives have a correlation $\rho = \sqrt{t}$ and variances $4/(tE_+)$ and $4/E_+$, respectively. Similarly, the observed effects for the test negatives can be simulated independently using a bivariate Normal distribution with correlation equal to $\rho = \sqrt{t}$ and variances $4/(tE_-)$ and

$4/E_-$, respectively. We simulated data representing 10,000 trial effects at the interim and final stages for threshold values of $TH = \{0.7, 0.8, 0.9\}$ and proportions of $t = \{20\%, 25\%, 33.3\%, 50\%\}$ events in the interim stage. We present analysis and results for $TH = 0.8$, which yields intermediate results and $\delta_* = \log(2/3)$. We give the results with other TH values in Appendix B for $t = 0.2, 0.5$.

We carried out prior specification by estimating for a grid of r_1, r_2, and p_{00} values, the proportion of trials in which the overall null H_{0+-} was rejected using simulations under all scenarios. The quantities r_1 and r_2 should be based on the *a priori* evidence for the classifier, but p_{00} should be set to ensure that type I errors for none of the null hypothesis H_{0+-}, H_{0+}, or H_{0-} exceeds 0.05. The values examined were $r_1, r_2 = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and sequence of p_{00} values equally spaced between 0 and 0.3. The action of rejecting a hypothesis was taken if the posterior probability was less than 0.05.

We chose an r_2 value of 0.1 for our priors as most appropriate for targeted therapy. Figure 1 gives the proportions of trials in which H_{0+-}, H_{0+}, or H_{0-} were rejected for a sequence of p_{00} values with $r_2 = 0.1$ using data simulated with $\delta_+ = \delta_* = 0$. These proportions are estimates of the type I errors. Among all r_1 values of interest, the probability of a type I error is less than 0.05 for values of p_{00} greater than or equal to 0.1. Subsequently, we chose a p_{00} value of 0.1 as this value provided greatest power. We therefore considered priors with the following parametrization:

1. P_1, with $p_{00} = 0.1, r_1 = 0.1, r_2 = 0.1$,
2. P_2, with $p_{00} = 0.1, r_1 = 0.5, r_2 = 0.1$, and
3. P_3, with $p_{00} = 0.1, r_1 = 0.9, r_2 = 0.1$.

The use of prior P_1 is appropriate in situations when one has low confidence in the classifier. Prior P_3 is for situations with high confidence in the classier, and prior P_2 provides a middle ground.

Using the trial design in Section 3, we estimated the following quantities for each scenario:

1. The proportion of trials in which H_{0-} is rejected at the end of the trial;
2. The proportion of trials in which H_{0+} is rejected at the end of the trial;
3. The proportion of trials in which the trial was stopped at the interim stage; and

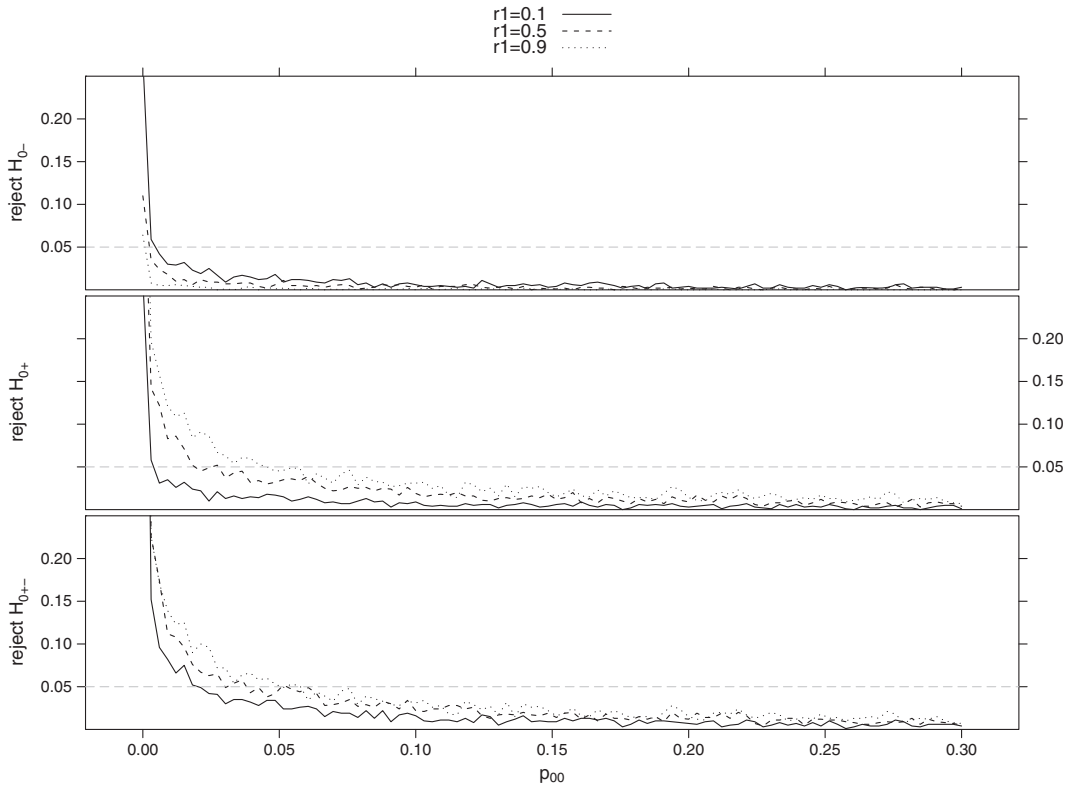

Figure 1. Estimated probability of rejecting H_{0+-} , H_{0+}, and H_{0-} versus p_{00} with $r_2 = 0.1$ for data simulated with $\delta_+ = \delta_- = 0$, $\delta_* = \log(2/3)$, and 50% prevalence.

4. The proportion of trials in which negative recruitment was stopped at the interim stage given that the trial was not stopped.

We took the action of rejecting the hypothesis H_{0-} if $P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) < 0.05$. Similarly, we took the action of rejecting the hypothesis H_{0+} if $P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) < 0.05$. We present the results for $\delta_* = \log(2/3)$ in Figures 2 and 3.

5. Results

The matrix plots given in Figures 2 and 3 summarize the results from the simulations. Figure 2 gives results for simulations with 25% prevalence, whereas Figure 3 for simulations with 50% prevalence. Rows 1–4 in the matrix plot represent the following respectively:

1. The proportion of trials in which H_{0-} was rejected;
2. The proportion of trials in which H_{0+} was rejected;
3. The proportion of trials in which accrual of marker-negative patients was stopped at the interim stage; and
4. The proportion of trials in which all accrual was stopped at the interim stage.

The columns in the matrix plot represent simulation scenarios with data generated on the basis of the following:

1. $\delta_+ = \delta_- = 0$,
2. $\delta_+ = \delta_*, \delta_- = 0$,
3. $\delta_+ = 0, \delta_- = \delta_*$, and
4. $\delta_+ = \delta_* = \delta_*$.

5.1. *Results with 25% prevalence of marker positivity*

Column 1 of Figure 2 shows results for the null simulation scenario in which $\delta_+ = \delta_- = 0$. Subplots of proportion of trials in which H_{0+} and H_{0-} were rejected show that type I error is well controlled

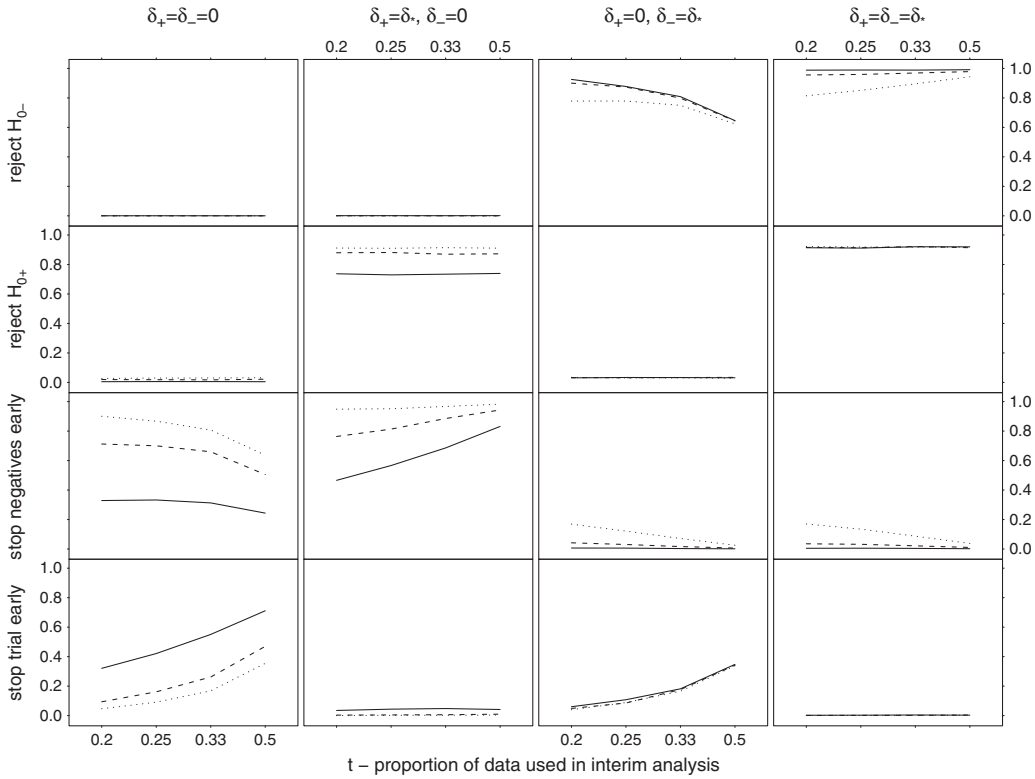

Figure 2. Matrix plot with results for simulations with 25% prevalence. Solid line represents P_1, dashed line P_2, and dotted line P_3.

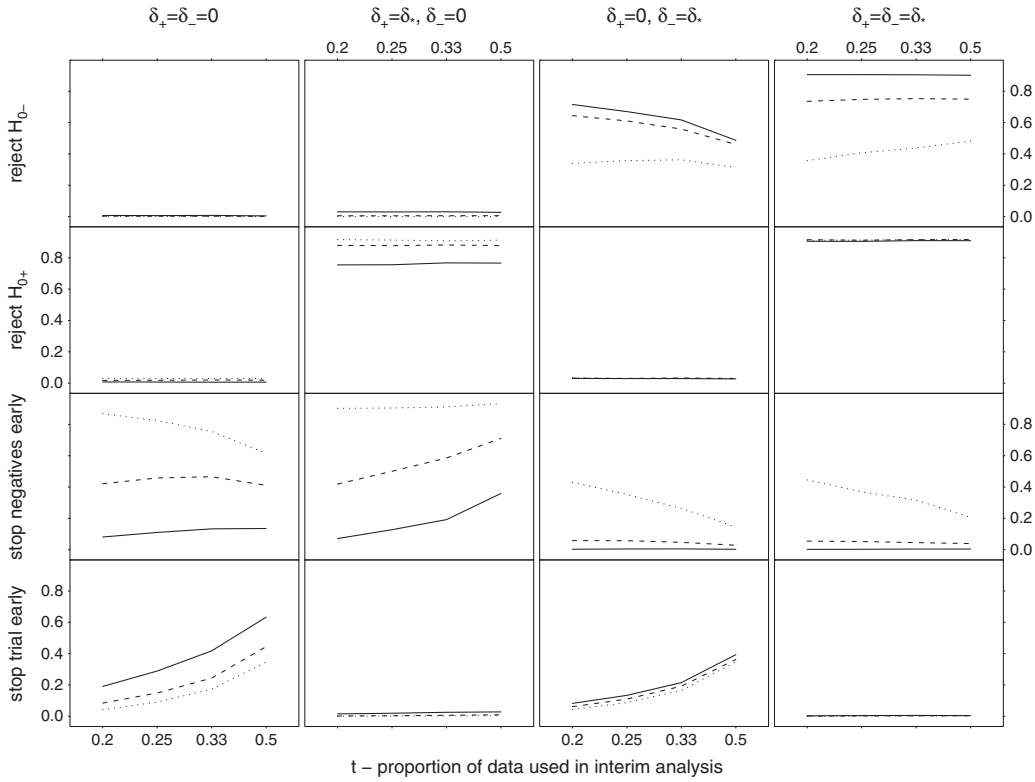

Figure 3. Matrix plot with results for simulations with 50% prevalence. Solid line represents P_1, dashed line P_2, and dotted line P_3.

with estimated values not exceeding 3.5% for all three priors and, at all times, interim analysis was performed. From a patient protection and cost perspective, a good design would have a high probability of stopping the trial at the interim stage. The highest proportions of trials are stopped at the interim stage under P_1, which is the prior that conveys a lack of confidence in the classifier results. The subplot also shows that the more data used in the interim stage, the higher the probability of stopping the trial under the global null; however, accrual of marker-negative patients is stopped early much less frequently when using prior P_1 compared with P_2 or P_3.

The second column of Figure 2 shows results for the scenario in which a treatment effect exists in the test positives only with $\delta_+ = \delta_*$, $\delta_- = 0$. The subplot on the proportion of trials in which H_{0-} was rejected indicates good control of type I error for H_{0-} for all three priors. The subplot on the proportional of trials in which H_{0+} was rejected indicates that all priors yield good power with values of at least 75%. Prior P_3 that conveys confidence in the classifier results in the highest power with values of at least 95%. For patient protection and cost considerations, an appropriate design for this scenario stops accrual of negatives at the interim stage without stopping the trial. The subplot on early termination of accrual of negatives indicates that the prior P_3 results in the highest probability of stopping accrual of negatives at the interim stage.

The third column in Figure 2 shows results for the unlikely scenario in which a treatment effect exist for the negatives only with $\delta_+ = 0$, $\delta_- = \delta_*$. The subplot of proportion of trials in which H_{0+} was rejected indicates good control of type I error for all priors with error estimates not exceeding 3.5%. The subplot of proportion of trials in which H_{0-} was rejected indicates that priors P_1 and P_2 provide the highest power estimates with values exceeding 90% for $t = 20\%$, with power decreasing, the later interim analysis is performed. This can be explained by the fact that the trial is terminated because of no treatment effect in positives with increasing frequency for later times of analysis. This reduces the power for rejecting H_{0-}.

The subplots in the last column of Figure 2 show results for the scenario in which a treatment effect exists for both the positives and the negatives, namely $\delta_+ = \delta_- = \delta_*$. The subplot of the proportion of trials in which H_{0-} was rejected shows that P_1 and P_2 provide very high power for rejecting H_{0-}

and the power for P_3 exceeds 80%. All priors provide high power for rejecting H_{0+} with values greater than 91%. The probability of stopping accrual of the negatives and stopping the trial at the interim stage for this scenario should be low, and all three priors result in low probabilities with estimated values not exceeding 0.17.

5.2. *Results with 50% prevalence of marker positivity*

The first column of Figure 3 shows results for the null scenario in which $\delta_+ = \delta_- = 0$. With 50% prevalence, we also observed good control of type I error for both H_{0+} and H_{0-} with values not exceeding 3.5%. Just as with 25% prevalence, the prior that best supports early termination is P_1, which is the prior that conveys lack of confidence in the classifier. The probabilities of early termination of marker-negative patients with priors P_2 and P_3 is reduced compared with that in Figure 2 because fewer marker-negative patients are available for analysis with 50% prevalence of marker positivity.

The second column of Figure 3 shows results for the scenario in which a treatment effect exists only in the test positives with $\delta_+ = \delta_*$, $\delta_- = 0$. With 50% prevalence, the results indicate good control of type I error for H_{0-} with all three priors. The power for rejecting H_{0+} is very similar to that for 25% prevalence. Table I compares proportion of trials in which accrual of negatives was stopped early for 25% versus 50% prevalence. From the table, it is clear that a futility analysis with P_3 stops negative recruitment in at least 90% of the trials regardless of the interim analysis time or prevalence. Using prior P_1 however, the proportion of trials in which accrual of marker negatives is stopped early is substantially reduced for 50% prevalence compared with 25% prevalence. This is also the case for prior P_2 but to a lesser degree. An important point from this scenario is that using a prior that conveys confidence in the classifier enables an early futility analysis without loss of power or increasing the type I error.

The subplots in the third column of Figure 3 show results for the unlikely scenario in which a treatment effect exists only in the negatives with $\delta_+ = 0, \delta_- = \delta_*$. Just as in the 25% prevalence analysis, type I error for H_{0+} is well controlled with values not exceeding 3.5% regardless of prior used and time at which interim analysis was performed. Power estimates for H_{0-} are comparatively less than those from 25% prevalence, the largest difference being with prior P_3. This can be attributed to the larger influence of the prior because of the smaller number of events in the test negatives at the interim and final stage. These phenomena also result in a higher probability of stopping accrual of negatives at the interim stage as well as lower probability of stopping the trial early.

The subplots in the last column in Figure 3 show the results for the scenario in which both the positives and negatives benefit from the treatment where $\delta_+ = \delta_- = \delta_*$. The results indicate good power for rejecting H_{0+} with values exceeding 90% for all three priors. Unlike the results for 25% prevalence, the performance of the three priors in the power for rejecting H_{0-} is markedly different; prior P_1 has high power estimates with values in the 90%–91% range for all t, prior P_2 has relatively good power with values in the 73%–75% range, whereas P_3 has power values not exceeding 50%. With the comparison of the plots of the proportion of trials in which H_{0-} is rejected for both 25% and 50% prevalence, the difference in power estimates are much larger for P_3 compared with P_2 and P_1. A reason for this difference is the larger sampling variation caused by fewer number of test-negative events. A good design for this scenario would have a low probability of stopping the trial as well as a low probability of stopping accrual of the negatives. Both P_1 and P_2 have low probability of stopping accrual of negatives early as well as probability of stopping the trial.

5.3. *Summary*

All three priors are completely adequate with regard to preserving all three type I errors under all conditions. They also provide adequate power for rejecting H_{0+} with P_2 and P_3 providing outstanding power.

Table I. Proportion of trials in which negative recruitment was stopped at interim stage for the simulation scenario with $\delta_+ = \delta_*$, $\delta_- = 0$.

	prev = 25%		prev = 50%	
Prior	$t = 0.2$	$t = 0.5$	$t = 0.2$	$t = 0.5$
P_1	0.47	0.83	0.07	0.36
P_2	0.76	0.94	0.42	0.71
P_3	0.95	0.98	0.90	0.93

Prior P_1 is substantially inferior to the others with regard to early stopping of accrual of marker-negative patients. Prior P_3 performs well for 25% prevalence but provides very poor power for rejecting H_{0-} when prevalence of positivity is 50%. At 25% prevalence, prior P_3 is superior to P_2 with regard to early stopping of accrual of marker-negative patients. When the prevalence is 25%, P_2 or P_3 is an appropriate prior. When prevalence is 50%, we recommend the use of the intermediate prior P_2.

6. Discussion

Our study demonstrates that confidence in a predictive classifier, based on biological information and early trial data, can be incorporated into the design of a randomized phase III clinical trial in a Bayesian framework in a manner that protects both patients and type I error. This design enables limiting the number of test-negative patients who are exposed to a drug that, based on biological evidence, seems unlikely to benefit them. This approach also limits costs of clinical development.

Another advantage of the Bayesian formulation is that it clarifies the nature of inference at the conclusion of the trial. Frequentist designs have been proposed that combine testing for an overall treatment effect with testing for a treatment effect in the test-positive subset, allocating the type I error between the two tests [20]. In practice, however, if the hypothesis of no overall treatment effect is rejected, and a primary biomarker has been measured, then investigators and regulators will likely insist on evaluating the treatment effects in the subsets. The Bayesian formulation formalizes that treatment effect in both subsets is ultimately of interest and utilizes the prior to determine how to share information across the strata. This leads to a very transparent and interpretable analysis.

Results from the simulations indicate that our two-stage Bayesian design provides adequate power for testing for an effect on the test positives. Most importantly, type I errors are well controlled. Type I error estimates for effects in both test positives and test negatives in the null simulation scenario were less than 0.035. Results in Section 5 indicate that even in the unlikely scenario where an effect exists for the marker negatives with no effect in the marker positives, empirical type I errors were less than 0.036.

With the appropriate prior, this Bayesian approach enables early futility analysis without substantial losses in power. The most relevant scenario in a futility analysis is one where there is a treatment effect in the test positives but no effect in the test negatives. Simulations using priors that reflect confidence in the biomarker resulted in 90% of the trials stopping negative accruals at the interim stage with as little as 20% of the planned event sizes. Ultimately, the decision on when to perform a futility analysis depends on the accrual rate of patients, the event rate, and the amount of follow-up time relative to survival time.

Different δ_* and prevalence of marker positivity change the marginal distribution of the data. For instance, low prevalence results in more marker-negative events and therefore less sample variation in the negative stratum, whereas smaller δ_* implies more marker-positive events and consequently more marker-negative events and less sample variation in both strata. Less sample variation results in less prior influence on the posterior. The operating characteristics of our method therefore depend on δ_* and the prevalence of marker positivity; hence, a careful selection of priors and evaluation of operating characteristics are needed for different δ_* and prevalence values.

The adaptive design of Wang *et al.* [7] makes a decision at the interim stage whether to maintain the stratified design or to only recruit test positives while maintaining the initially planned total sample size. Hence, the number of test-positive patients accrued at the end is much greater if accrual of test negatives is terminated at the interim stage. Our two-stage design has three main advantages over the design of Wang *et al.* Our adaptive design incorporates a degree of prior confidence in the biomarker utility at the interim stage to make informed decisions affecting patient accrual, whereas that of Wang *et al.* assumes complete confidence in the biomarker. The approach of Wang *et al.* does not consider the possibility of early termination of the entire trial, whereas our approach makes a decision on early termination of the trial depending on the efficacy of the treatment in the test positives at the interim stage. The posterior distribution of treatment effect for our design yielded high power (greater than 90% with the prior that reflects confidence in the classifier) for rejecting H_{0+} with 50% prevalence in the simulation scenario, where a treatment effect exists for the test positives and no treatment effect exists in the test negatives. The adaptive design of Wang *et al.* also yields high power for the positives (greater than 90%) but with a much greater expected sample size of test-positive patients. This results in large increases in the number of patients screened, thereby prolonging the trial. This effect becomes even greater if the prevalence of the marker positives is lower.

There are several extensions that can be studied within the framework provided here. We initially size the number of events in marker-positive patients to detect a treatment effect in the marker positives on

the basis of a prespecified power value for a specified δ_*. We used this number to obtain the expected number of test-negative events assuming a nonprognostic classifier; hence, the power for rejecting H_{0-} is not directly controlled. Other approaches are possible.

Our approach could be extended to more than two strata where stratification is based on monotonic level of marker positivity. In addition to the parameter p_{00}, parameters for all marginal conditional null probabilities such as those specified in Equations (4) and (5), as well as all combinations of joint conditional null probabilities, need to be specified so as to define the prior's probability mass function. For example, with three strata labeled 1, 2, and 3 where stratum 3 has the highest level of marker positivity and stratum 1 has the lowest, if treatment effects in the strata are denoted as δ_1, δ_2, and δ_3, the prior's parametrization is as follows:

$$P(\delta_1 = 0|\delta_2 = \delta_3 = \delta_*) = r_{1|23},$$
$$P(\delta_2 = 0|\delta_1 = \delta_3 = \delta_*) = r_{2|13},$$
$$P(\delta_3 = 0|\delta_1 = \delta_2 = \delta_*) = r_{3|12},$$

and

$$P(\delta_1 = \delta_2 = 0|\delta_3 = \delta_*) = r_{12|3},$$
$$P(\delta_1 = \delta_3 = 0|\delta_2 = \delta_*) = r_{13|2},$$
$$P(\delta_2 = \delta_3 = 0|\delta_1 = \delta_*) = r_{23|1},$$

where $P(\delta_1 = \delta_2 = \delta_3 = 0) = p_{00}$. One strategy in the trial design at the interim stage is to test the effectiveness of the treatment in patients with high marker positivity first. For the three-strata example, if the treatment is deemed ineffective in stratum 3, then the trial is stopped. Otherwise, if the treatment is deemed ineffective in stratum 2, then the trial is continued with accrual of stratum 3 patients only, and so on. The numbers of parameters can be reduced on the basis of monotonicity. In the example, it is reasonable to assume $r_{2|13} = r_{13|2} = 0$. This approach works well with a small number of strata, but for many strata or when strata are not ordered, other approaches need to be considered, for example, using continuous distributions for the prior.

Our methodology can be applied to dose-selection trials where an efficacy endpoint is assumed to have a monotonic relationship with dose. For this application, the trial population should be stratified by the treatment arms that correspond to different dose levels. For example, in a trial with two dose levels, high and low, prior information on the efficacy of the dose levels can be incorporated into the design by selecting appropriate values of r_1 and r_2, where r_1 represents the probability of the treatment being ineffective at the higher dose given that it works in the lower dose and r_2 represents the probability of the treatment being ineffective at the lower dose given it works in the higher dose. The interim analysis outlined in Section 3 is directly applicable, likening the higher dose's treatment arm with the marker-positive strata and the lower dose's treatment arm with the marker-negative strata. Hence, if the treatment is deemed ineffective at the higher dosage, then the trial should be stopped at the interim stage. Incorporation of prior belief in efficacy of dose levels in trial design has an advantage over the practice of treating different doses as different treatments [21] as it allows for information sharing among the treatment arms.

For our simulations, the prior distribution for (δ_+, δ_-) have mass on $\{(0,0), (\delta_*, 0), (0, \delta_*), (\delta_*, \delta_*)\}$, where δ_* represents the treatment effect of minimal clinical significance. Frequentist designs are typically planned on the basis of such discretizations. This discretization provides a foundation for the continuous priors. It would be useful to evaluate in the future the advantages offered by such continuous priors, but the design based on the discretized space can be useful in designing new studies.

In the design proposed, we reject a hypothesis if its posterior probability falls below a prespecified level of $\epsilon = 0.05$. At the interim stage, we do not reject a hypothesis if the posterior is greater than some threshold value $TH = 0.7$, 0.8, or 0.9. The designs with the parameters considered worked well for all simulation scenarios examined, but other values could be examined for achieving acceptable operating characteristics in other conditions.

This paper outlines a design that is very promising in developing molecularly targeted treatments. The framework set out in this paper offers a practical approach for utilizing Bayesian designs in phase III trials. The key is not computational power but careful selection of prior distributions that are appropriate for the context of the study and ensuring that the design has good frequency characteristics under a range of values for δ_+ and δ_-. This approach enables one to utilize prior biological evidence while retaining the strengths of the frequentist formulation that has been so valuable for clinical trials.

APPENDIX A. Posterior probabilities calculation

The prior parametrization with

$$P(\delta_+ = 0|\delta_- = \delta_*) = r_2$$
$$P(\delta_- = 0|\delta_+ = \delta_*) = r_1$$
$$P(\delta_+ = 0, \delta_- = 0) = p_{00}$$

results in the following probability mass function:

$$P(\delta_+ = 0, \delta_- = \delta_*) = aq_2(1 - p_{00})$$
$$P(\delta_+ = \delta_*, \delta_- = 0) = aq_1(1 - p_{00})$$
$$P(\delta_+ = \delta_*, \delta_- = \delta_*) = a(1 - p_{00})$$
$$P(\delta_+ = 0, \delta_- = \delta_*) = p_{00},$$

where $a = (1 - r_1)(1 - r_2)/(1 - r_1 r_2)$, $q_1 = r_1/(1 - r_1)$, $q_2 = r_2/(1 - r_2)$, and $q_{00} = p_{00}/(1 - p_{00})$. We let $L_{i,j} = P(\hat{\delta}_+, \hat{\delta}_-|\delta_+ = i, \delta_- = j)$ denote the marginal conditional probabilities of the bivariate Normal density given in Equation (3), then the posterior probabilities are as follows:

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) = (p_{00}L_{0,0} + P(\delta_+ = 0, \delta_- = \delta_*)L_{0,\delta_*})/k, \tag{10}$$

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = (p_{00}L_{0,0} + P(\delta_+ = \delta_*, \delta_- = 0)L_{\delta_*,0})/k, \tag{11}$$

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = (p_{00}L_{0,0})/k, \tag{12}$$

where

$$k = p_{00}L_{0,0} + P(\delta_+ = \delta_*, \delta_- = 0)L_{\delta_*,0} + P(\delta_+ = 0, \delta_- = \delta_*)L_{0,\delta_*} + P(\delta_+ = \delta_*, \delta_- = \delta_*)L_{\delta_*,\delta_*}. \tag{13}$$

By factoring out the numerator in both the numerator and denominator of Equations (10) and (12), the posterior probabilities can be rewritten as

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_1L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0} + aq_2L_{0,\delta_*}}\right)^{-1}, \tag{14}$$

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2L_{0,\delta_*} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0} + aq_1L_{\delta_*,0}}\right)^{-1}, \tag{15}$$

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) = \left(1 + \frac{aq_2L_{0,\delta_*} + aq_1L_{\delta_*,0} + aL_{\delta_*,\delta_*}}{q_{00}L_{0,0}}\right)^{-1}. \tag{16}$$

A1. Limiting probabilities

Complete *a priori* confidence in the utility of the classifier means $r_1 \to 1$ and $r_2 \to 0$ then $a \to 0$ and

$$aq_1 = r_1(1 - r_2)/(1 - r_1 r_2) \to 1,$$

$$aq_2 = r_2(1 - r_1)/(1 - r_1 r_2) \to 0,$$

and the limiting posterior probability of the treatment having no effect in the marker positive patients is

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) \to \left(1 + \frac{L_{\delta_*,0}}{q_{00}L_{0,0}}\right)^{-1}, \tag{17}$$

which is dependent on p_{00} and the data. The limiting posterior probability of the treatment having no effect in the marker-negative patients is

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) \to 1. \tag{18}$$

The limiting joint posterior probability of the treatment having no effect in both marker-positive and marker-negative patients is

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) \to \left(1 + \frac{L_{\delta_*,0}}{q_{00}L_{0,0}}\right)^{-1}. \tag{19}$$

A complete lack of confidence in the utility of the biomarker implies $r_1 \to 0$ and $r_2 \to 0$ then $a \to 1$. As a result

$$aq_1 \to 0,$$

$$aq_2 \to 0,$$

and

$$P(\delta_+ = 0|\hat{\delta}_+, \hat{\delta}_-) \to \left(1 + \frac{L_{\delta_*,\delta_*}}{q_{00}L_{0,0}}\right)^{-1}, \tag{20}$$

$$P(\delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) \to \left(1 + \frac{L_{\delta_*,\delta_*}}{q_{00}L_{0,0}}\right)^{-1}, \tag{21}$$

$$P(\delta_+ = 0, \delta_- = 0|\hat{\delta}_+, \hat{\delta}_-) \to \left(1 + \frac{L_{\delta_*,\delta_*}}{q_{00}L_{0,0}}\right)^{-1}. \tag{22}$$

APPENDIX B. Interim analysis results

Table BI. Simulation results for 25% prevalence showing probability of stopping accrual of negatives early (S1), probability of stopping trial early (S2), probability of rejecting H_{0+} (PR+), and probability of rejecting H_{0-} (PR−) with 20% and 50% accrued events for interim analysis.

			Prior $P1$				Prior $P2$				Prior $P3$			
(δ_+, δ_-)	t	TH	S1	S2	PR+	PR−	S1	S2	PR+	PR−	S1	S2	PR+	PR−
(0, 0)	0.2	0.7	0.28	0.46	0.00	0.00	0.69	0.17	0.02	0.00	0.88	0.09	0.03	0.00
		0.8	0.33	0.32	0.00	0.00	0.71	0.09	0.02	0.00	0.90	0.05	0.03	0.00
		0.9	0.35	0.17	0.00	0.00	0.67	0.03	0.02	0.00	0.89	0.01	0.03	0.00
	0.5	0.7	0.18	0.79	0.00	0.00	0.43	0.55	0.02	0.00	0.54	0.45	0.03	0.00
		0.8	0.24	0.71	0.00	0.00	0.51	0.47	0.02	0.00	0.64	0.35	0.03	0.00
		0.9	0.35	0.58	0.00	0.00	0.63	0.32	0.02	0.00	0.75	0.24	0.03	0.00
$(\delta_*, 0)$	0.2	0.7	0.53	0.07	0.72	0.00	0.82	0.01	0.87	0.00	0.96	0.00	0.91	0.00
		0.8	0.47	0.03	0.74	0.00	0.76	0.00	0.88	0.00	0.95	0.00	0.91	0.00
		0.9	0.37	0.01	0.74	0.00	0.66	0.00	0.88	0.00	0.89	0.00	0.91	0.00
	0.5	0.7	0.83	0.07	0.74	0.00	0.95	0.02	0.87	0.00	0.98	0.01	0.91	0.00
		0.8	0.83	0.04	0.74	0.00	0.94	0.01	0.87	0.00	0.98	0.00	0.91	0.00
		0.9	0.81	0.02	0.74	0.00	0.92	0.00	0.88	0.00	0.98	0.00	0.91	0.00
$(0, \delta_*)$	0.2	0.7	0.01	0.12	0.03	0.87	0.05	0.10	0.03	0.84	0.22	0.09	0.03	0.68
		0.8	0.01	0.06	0.03	0.93	0.04	0.05	0.03	0.90	0.17	0.04	0.03	0.78
		0.9	0.00	0.02	0.03	0.97	0.02	0.01	0.03	0.96	0.10	0.01	0.03	0.87
	0.5	0.7	0.00	0.44	0.03	0.56	0.01	0.44	0.04	0.55	0.03	0.42	0.03	0.54
		0.8	0.00	0.35	0.03	0.65	0.01	0.34	0.03	0.64	0.03	0.34	0.03	0.62
		0.9	0.00	0.22	0.03	0.77	0.00	0.22	0.03	0.77	0.02	0.23	0.03	0.73
(δ_*, δ_*)	0.2	0.7	0.01	0.01	0.92	0.98	0.06	0.00	0.92	0.93	0.24	0.00	0.92	0.74
		0.8	0.00	0.00	0.91	0.99	0.04	0.00	0.92	0.96	0.17	0.00	0.92	0.81
		0.9	0.00	0.00	0.92	0.99	0.02	0.00	0.92	0.97	0.10	0.00	0.92	0.88
	0.5	0.7	0.00	0.01	0.92	0.99	0.01	0.01	0.92	0.97	0.05	0.01	0.92	0.92
		0.8	0.00	0.00	0.92	0.99	0.01	0.00	0.91	0.98	0.04	0.00	0.92	0.94
		0.9	0.00	0.00	0.92	1.00	0.01	0.00	0.92	0.98	0.03	0.00	0.92	0.96

Table BII. Simulation results for 50% prevalence showing probability of stopping accrual of negatives early (S1), probability of stopping trial early (S2), probability of rejecting H_{0+} (PR+), and probability of rejecting H_{0-} (PR−) with 20% and 50% accrued events for interim analysis.

			Prior $P1$				Prior $P2$				Prior $P3$			
(δ_+, δ_-)	t	TH	S1	S2	PR+	PR−	S1	S2	PR+	PR−	S1	S2	PR+	PR−
$(0, 0)$	0.2	0.7	0.10	0.30	0.01	0.01	0.50	0.16	0.02	0.00	0.86	0.09	0.03	0.00
		0.8	0.08	0.19	0.01	0.01	0.42	0.08	0.02	0.00	0.87	0.04	0.03	0.00
		0.9	0.05	0.09	0.01	0.01	0.25	0.02	0.02	0.00	0.77	0.01	0.03	0.00
	0.5	0.7	0.12	0.71	0.01	0.00	0.36	0.54	0.02	0.00	0.52	0.45	0.03	0.00
		0.8	0.14	0.63	0.01	0.00	0.41	0.44	0.02	0.00	0.62	0.35	0.03	0.00
		0.9	0.15	0.50	0.01	0.01	0.42	0.32	0.02	0.00	0.67	0.23	0.03	0.00
$(\delta_*, 0)$	0.2	0.7	0.13	0.03	0.76	0.03	0.56	0.01	0.88	0.00	0.95	0.00	0.91	0.00
		0.8	0.07	0.02	0.75	0.03	0.42	0.00	0.88	0.01	0.90	0.00	0.92	0.00
		0.9	0.03	0.00	0.76	0.03	0.22	0.00	0.88	0.00	0.77	0.00	0.91	0.00
	0.5	0.7	0.43	0.05	0.77	0.03	0.77	0.01	0.88	0.01	0.95	0.01	0.91	0.00
		0.8	0.36	0.03	0.77	0.03	0.71	0.01	0.88	0.01	0.93	0.00	0.91	0.00
		0.9	0.25	0.01	0.77	0.03	0.57	0.00	0.88	0.01	0.87	0.00	0.91	0.00
$(0, \delta_*)$	0.2	0.7	0.01	0.15	0.03	0.66	0.10	0.12	0.03	0.59	0.55	0.09	0.03	0.24
		0.8	0.00	0.08	0.03	0.72	0.06	0.06	0.03	0.64	0.43	0.04	0.03	0.34
		0.9	0.00	0.03	0.03	0.76	0.02	0.02	0.03	0.69	0.24	0.01	0.03	0.43
	0.5	0.7	0.01	0.47	0.03	0.43	0.04	0.46	0.03	0.38	0.17	0.44	0.03	0.26
		0.8	0.00	0.39	0.03	0.49	0.03	0.36	0.03	0.46	0.14	0.35	0.03	0.31
		0.9	0.00	0.26	0.03	0.60	0.02	0.24	0.03	0.54	0.10	0.23	0.03	0.38
(δ_*, δ_*)	0.2	0.7	0.01	0.01	0.91	0.90	0.11	0.00	0.91	0.70	0.59	0.00	0.91	0.29
		0.8	0.00	0.00	0.91	0.91	0.05	0.00	0.92	0.74	0.45	0.00	0.91	0.36
		0.9	0.00	0.00	0.91	0.91	0.01	0.00	0.92	0.75	0.25	0.00	0.92	0.44
	0.5	0.7	0.01	0.01	0.91	0.90	0.07	0.01	0.91	0.74	0.29	0.01	0.92	0.46
		0.8	0.00	0.00	0.91	0.90	0.04	0.00	0.92	0.75	0.21	0.00	0.92	0.48
		0.9	0.00	0.00	0.91	0.91	0.02	0.00	0.91	0.75	0.13	0.00	0.92	0.50

References

1. Simon R. Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert Opinion in Medical Diagnostics* 2008; **2**:721–729.
2. Temple RJ. Special study designs: early escape, enrichment, studies in non-responders. *Communications in Statistics - Theory and Methods* 1994; **23**(2):499–531. DOI: 10.1080/03610929408831269.
3. Maitournam A, Simon R. On the efficency of targeted clinical trials. *Statistics in Medicine* 2005; **24**:329–339. DOI: 10.1002/sim.1975.
4. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 2004; **10**:6759–6763. DOI: 10.1158/1078-0432.CCR-04-0496.
5. Simon R. The use of genomics in clinical trial design. *Clinical Cancer Research* 2008; **14**:5984–5993. DOI: 10.1158/1078-0432.CCR-07-4531.
6. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* 2005; **23**(9):2020–2027. DOI: 10.1200/JCO.2005.01.112.
7. Wang S, O'Neill RT, Hung JMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**:227–244. DOI: 10.1002/pst.300.
8. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 2005; **11**:7872–7878.
9. Jiang W, Freidlin B, Simon R. Biomarker adaptive threshold design: treatment with possible biomarker-defined subset effect. *Journal on National Cancer Institute* 2007; **99**:1036–1043. DOI: 10.1093/jnci/djm022.
10. Berry DA. A case for Bayesianism in clinical trials. *Statistics in Medicine* 1993; **12**:1377–1393. DOI: 10.1002/sim.4780121504.
11. Greenhouse JB, Wasserman L. Robust bayesian methods for monitoring clinical trials. *Statistics in Medicine* 1995; **14**:1379–1391. DOI: 10.1002/sim.4780141210.
12. Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* 1993; **42**:341–353.
13. Carlin BP, Sargent DJ. Robust bayesian approaches for clinical trial monitoring. *Statistics in Medicine* 1996; **15**:1093–1106. DOI: 10.1002/(SICI)1097-0258(19960615)15:11<1093::AID-SIM231>3.0.CO;2-0.
14. Giaccone G, Herbst RS, Manegold C, Scagliotti G, Rosell R, Miller V, Natale RB, Schiller JH, von Pawel J, Pluzanska A, et al. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small-cell lung cancer: a phase III trial–INTACT 1. *Journal of Clinical Oncology* 2004; **22**(5):777–784. DOI: 10.1200/JCO.2004.08.001.

15. Van Cutsem E, Peeters M, Siena S, Humblet Y, Hendlisz A, Neyns B, Canon JL, Van Laethem JL, Maurel J, Richardson G, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. *Journal of Clinical Oncology* 2007; **25**(13):1658–1664. DOI: 10.1200/JCO.2006.08.1620.
16. Moore MJ, Goldstein D, Hamm J, Figer A, Hecht JR, Gallinger S, Au HJ, Murawa P, Walde D, Wolff RA, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *Journal of Clinical Oncology* 2007; **25**(15):1960–1966. DOI: 10.1200/JCO.2006.07.9525.
17. Ozols RF, Bundy BN, Greer BE, Fowler JM, Clarke-Pearson D, Burger RA, Mannel RS, DeGeest K, Hartenbach EM, Baergen R. Phase III trial of carboplatin and paclitaxel compared with cisplatin and paclitaxel in patients with optimally resected stage III ovarian cancer: a Gynecologic Oncology Group study. *Journal of Clinical Oncology* 2003; **21**(17):3194–3200. DOI: 10.1200/JCO.2003.02.153.
18. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, Gianni L, Baselga J, Bell R, Jackisch C, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *New England Journal of Medicine* 2005; **353**(16):1659–1672. DOI: 10.1056/NEJMoa052306.
19. Schaid DJ, Wieand S, Therneau TM. Optimal two-stage screening designs for survival comparison. *Biometrika* 1990; **7**:507–513.
20. Simon R, Wang S. Use of genomic signatures in therapeutics development in oncology and other diseases. *The Pharmacogenomic Journal* 2006; **6**:166–173.
21. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703. DOI: 10.1002/sim.1362.