Published in final edited form as:

Biometrics. 2008 September ; 64(3): 886–894. doi:10.1111/j.1541-0420.2007.00951.x.

Bayesian Optimal Design for Phase II Screening Trials

Meichun Ding¹, Gary L. Rosner^{2,*}, and Peter Müller²

¹Hoffman-La Roche Inc., Department of Biostatistics, MS 44, 340 Kinsland Street, Nutley, NJ 07110-1199

²Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030

SUMMARY

Most phase II screening designs available in the literature consider one treatment at a time. Each study is considered in isolation. We propose a more systematic decision-making approach to the phase II screening process. The sequential design allows for more efficiency and greater learning about treatments. The approach incorporates a Bayesian hierarchical model that allows combining information across several related studies in a formal way and improves estimation in small data sets by borrowing strength from other treatments. The design incorporates a utility function that includes sampling costs and possible future payoff. Computer simulations show that this method has high probability of discarding treatments with low success rates and moving treatments with high success rates to phase III trial.

Keywords

Backward induction; Bayesian; Decision-theoretic; Phase II screening trials

1. Introduction

A number of frequentist and Bayesian proposals for phase II clinical study designs have appeared over the last three decades. These papers typically consider a binary outcome for which we use the generic terms “success” and “failure.” Gehan (1961) and Simon (1989) proposed popular 2-stage frequentist designs that fix the total sample size.

There are essentially two Bayesian approaches to single-treatment studies, such as phase II studies in oncology. One is a decision-theoretic approach, e.g., Berry and Pearson (1985), Sylvester (1988), and Stallard, Thall, and Whitehead (2000). These designs seek to maximize the expected value of a pre-specified utility function. The second Bayesian approach does not explicitly specify a loss function. Instead, stopping criteria relate to whether the posterior probability of some clinically important event exceeds a threshold. Thall and Simon’s (1994) design is an example that has led to many extensions. Heitjian (1997) proposed stopping rules based on posterior probabilities computed using two different prior distributions.

Yao, Begg, and Livingston (1996) considered screening of new treatments as a continuous process, namely, a sequence of clinical studies that ends once a promising treatment appears.

*grosner@mdanderson.org .

SUPPLEMENTARY MATERIALS Web Appendix referenced in Section 3 is available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

The sample size in each study is fixed, but the total sample size (or the number of studies) until identification of a promising agent is not. Their design aims to minimize the expected sample size until a promising treatment is identified. In later work, Yao and Venkatraman (1998), Wang and Leung (1998), and Leung and Wang (2001) considered a variety of extensions leading to two-stage designs and fully sequential designs. Taking a Bayesian decision-theoretic approach, Rossell, Müller, and Rosner (2007) find optimal linear boundaries for fully sequential phase II screening studies. Their boundaries are optimal with respect to a utility function that incorporates sampling cost and potential pay-off if the pivotal trial shows significant benefit, similar to utility function at the end of Section 2.3. Stout and Hardwick (2005) present a general decision-theoretic framework for finding optimal designs for screening trials, incorporating costs and constraints.

Many designs assume that the prior distribution for the treatment-specific success probabilities is a uniform Beta (1,1) distribution. In essence, these designs are based on classical hypothesis testing and are driven by the desire to reduce the risk of making incorrect decisions with respect to statistical hypothesis tests. In other words, these proposals seek to minimize frequentist Type I and Type II error probabilities.

Whether one considers a frequentist or a Bayesian approach, most screening design proposals in the oncology literature consider one treatment at a time. Each study exists in isolation without benefit of learning from past treatments. The designs do not formally incorporate updating prior knowledge about the success probability of the next treatment under consideration.

In this paper, we take a Bayesian decision-theoretic approach to design single-arm phase II screening trials. Our approach allows combining information across several related studies in a formal way. Furthermore, the design improves estimation in small data sets by borrowing strength from other studies and/or treatments through a Bayesian hierarchical structure. Additionally, our design considers the reasons for carrying out the studies and their goals through an appropriate utility function. One can use whatever utility function makes sense for the particular application. Here, we consider a utility function that incorporates costs and financial rewards arising from the drug development perspective. Specifically, there are sampling cost per patient for data collection. Additionally, there will be a gain if the new treatment shows significant benefit in a future phase III randomized clinical trial.

Briefly, our approach allows one to produce a decision table that one can use for all treatments or studies. The decision table relates the optimal decision at each interim analysis to a low-level summary of the posterior distribution, the posterior mean. By incorporating Bayesian updating in a hierarchical model and basing decisions on the posterior mean, our design approach allows decisions about further development of new treatments to account for learning about the treatments that have already gone through the screening process.

In the next section we describe the probabilistic setup of our model and the utility function we use. We then discuss the determination of decision tables in Section 3. In Section 4, we compare our method to the fully sequential design of Leung and Wang (2001) and present some simulation-based results relating to sensitivity of our method to prior specification. We conclude with a discussion in Section 5.

2. Design

2.1 Setup

We consider a sequential decision problem in which later decisions depend on earlier outcomes. We assume a dichotomous clinical endpoint and an indefinite sequence of new treatments for testing. The subscript t refers to the time of an analysis-decision. For us, time starts with the first treatment or study and continues until one identifies a treatment for phase III evaluation. At analysis-decision time t , patient j on treatment or study i gives rise to a binomial random variable y_{tij} with unknown success probability θ_i . Study i may concern a new drug, a new combination of drugs, or a new indication for an existing drug, for example, and we use “study” and “treatment” interchangeably. We assume that one will observe the response soon enough to enable the sequential designs. Also, we assume that the same maximum number of interim analyses applies to each study.

At each analysis-decision time t , we choose a decision d_{ti} for study i . Decision $d_{ti} = 1$ indicates “stop study i and discard the experimental treatment”; $d_{ti} = 2$ indicates “stop study i and switch to phase III trial”; decision $d_{ti} = 3$ indicates “continue entering patients into study i until the next analysis-decision time $(t+1)$ ”. Let $D_t = (d_1, \dots, d_t)$ be the set of decisions for all of the studies up to time t . We let Y_t be all the data observed for all the patients and studies up to time t . The history up to time t is $H_t = (D_{t-1}, Y_t)$, consisting of all the decisions up to time $t - 1$ and all the data up to time t . Decision d_{ti} can depend on all the data up to time t and all the decisions up to time $t - 1$, i.e., $d_{ti} = d_{ti}(H_t)$.

Each study is a single-arm design without treatment comparisons. One reason we consider a single-arm design is that classical two-arm trials often require larger sample sizes than are practical for phase II studies. The goal of many phase II trials in oncology is to screen new treatments for activity, unlike phase III confirmatory trials. Phase III clinical trials typically seek to provide definitive evidence of a treatment’s clinical effectiveness (superiority or noninferiority) relative to a placebo or standard treatment. The goal of phase II trials does not imply an explicit comparison with existing treatments, *per se*. That said, phase II studies implicitly have a comparative aspect that treatments have to show promise in order to pass the screening process and go to a phase III trial.

2.2 Hierarchical Model

We assume a three-stage hierarchical model. At the first stage of the hierarchy, we model the number of successes for each treatment at each analysis-decision time. The outcome data for patient j on treatment i at analysis-decision time t is binomially distributed, $y_{tij} | \theta_i \sim \text{Bin}(1, \theta_i)$. At the second level, we specify a population distribution for the success probability of each treatment or in each study, θ_i . We assume a probit regression model $\theta_i = \Phi(\mu_i)$, where $\Phi()$ denotes the standard normal cumulative distribution function. We characterize our uncertainty about the distribution of treatment-specific success probabilities via μ_i . We consider that the probit model parameters (μ_i , corresponding to the treatment-specific success probabilities θ_i) are normally distributed with mean θ_0 . At the bottom level, uncertainty about the hyperparameter θ_0 is also characterized by a normal distribution. The hierarchical probability model is:

$$\begin{aligned} y_{tij} | \theta_i &\sim \text{Bin}(1, \theta_i) \\ \theta_i &= \Phi(\mu_i), \mu_i | \mu_0 \sim N(\mu_0, \sigma^2) \\ \mu_0 &\sim N(\gamma, \tau^2), \quad \text{with } \gamma, \sigma^2, \text{ \& } \tau^2 \text{ known} \end{aligned} \tag{1}$$

The first equation in the hierarchical probit model represents the likelihood. The last two equations represent the prior distribution of the study-specific success probabilities.

Inference focuses on μ_0 and μ_i . The posterior distribution of μ_i , and, therefore θ_i , a deterministic function of μ_i , characterizes current knowledge about the success probability of treatment or study i . Since all treatments are connected to each other through the hyperprior normal distribution with parameter μ_0 , we continually update our knowledge about μ_0 in light of data generated by the current and previous studies. The marginal posterior distribution of μ_0 based on the first $i - 1$ studies can be considered as the prior for the i -th study. In this way, we combine information across studies and treatments in a formal probabilistic way and improve estimation in small datasets by borrowing information from other treatments.

We work with a mathematically equivalent statement of the probit model, one that is more amenable to the implementation of stochastic posterior simulation, as in Albert and Chib (1993). We introduce latent variables, viz., z_{tij} for patient j on treatment i at analysis-decision time t . If $z_{tij} | \mu_i \sim N(\mu_i, 1)$ and $y_{tij} = \{1 \text{ if } z_{tij} \geq 0; 0 \text{ otherwise}\}$, then $P(y_{tij} = 1) = P(z_{tij} \geq 0) = \Phi(\mu_i) = \theta_i$.

2.3 Utility Function

We consider a utility function on the basis of financial costs and potential gains from the drug developer's perspective. We assume a fixed sampling cost c_1 per patient in the phase II study. The fixed number of patients recruited between any two analysis-decision times in the phase II study is n_1 . With a maximum of T decision time points in each phase II study, the maximum number of patients is $T \times n_1$ per study. For the subsequent phase III trial, we let c_2 be the cost per patient. Phase III clinical trials are usually more expensive, because they typically involve many hospitals, they often require longer follow-up, and regulatory agencies may require more extensive data collection.

The gain in the utility function comes from the degree of benefit shown in the phase III study for the new treatment. Let the random variable Δ be the difference of success probabilities (new treatment minus the old one). We characterize the reward for improving the success probability of the new treatment over the old or current standard treatment as being proportional to Δ , with a benefit of b per unit of Δ . That is, if a pivotal phase III trial concludes that the drug is, in fact, an effective treatment, there is a positive benefit $b * \Delta$. If, on the other hand, the pivotal trial claims that the treatment is inefficacious, then there is no benefit at all. Mathematically, the benefit can be written as $\max(0, b * \Delta) \times I_{\{SIG\}}$, with the indicator function $I_{\{SIG\}} = 1$ if the phase III study concludes with a statistically significant outcome in favor of the new treatment and 0 otherwise. The phase III benefit in the utility calculation is a probabilistic calculation involving the posterior predictive probability of success for the new treatment in phase III, given the phase II data.

Let Y_t denote the data observed up to time t from the current study and all earlier phase II studies. We let Y_{III} represent the future phase III trial data, with n_2 patients included in the phase III trial. We follow current practice for determining the sample size for a phase III clinical trial in a frequentist hypothesis-testing framework. That is, we consider a one-sided, level 0.05 significance test of no difference with 90% power for the alternative $\delta = (\theta_{new} - \theta_{old}) > 0$. In our example, the success probability of the standard or control treatment, θ_{old} , is 0.2, and the difference, δ , is 0.3 for the purpose of illustration. The future phase III trial will randomize n_2 patients equally to the new treatment and the standard one. We do not consider sequential sampling in the phase III trial, although one could easily incorporate it via posterior predictive sampling of the future clinical trial.

For the purpose of specifying the utility function, we only need a stylized description of the phase III trial, i.e., a description that captures the important features of the implication of a stopping decision in the current trial. The utility depends on the outcome of the phase III trial, which is not yet known. We compute the expected utility by integrating with respect to the predictive distribution of the phase III data. The predictive distribution for the future phase III study integrates over the posterior distribution for the current treatment and the distribution of the success probability of the control treatment. In our example, we characterize the variation of the success probability of the standard or control treatment, θ_{old} , with a Beta (20, 80) distribution, having mean 0.2, variance 0.002, and 95% central interval (0.128, 0.283). In general, one would want to allow for relative certainty about the standard or control treatment's success probability in the future study and not set the variance too large.

The test statistic for the phase III trial is $Z = (\hat{p}_{new} - \hat{p}_{old}) / \sqrt{\hat{v}_{new} + \hat{v}_{old}}$. Here, $\hat{p}_{new}$ and $\hat{p}_{old}$ are the observed proportions of successes for the new and standard treatments, respectively, the usual frequentist estimates of the success probabilities θ_{new} and θ_{old} . In the denominator, $\hat{v}_k = 2\hat{p}_k(1 - \hat{p}_k)/n_k$, $k = new$ or old , is the variance of the estimator of the treatment-specific success probability. The standard one-sided test calls the new treatment efficacious if $z > z_{1-\alpha}$.

The utility function at time t , $u_t(d_t, \theta_{new}, \theta_{old}, Y_t, Y_{III})$, equals $-c_1 * n_1 * t$ if the decision is to stop and discard the treatment. If the decision is to switch to phase III, then the utility equals $-\{c_1 * n_1 * t + c_2 * n_2\} + b * \{\theta_{new} - \theta_{old}\} I_{[z > z_{1-\alpha}]}$.

3. Decision Tables

The traditional solution to optimal sequential design problem is full backward induction, also known as dynamic programming (DeGroot, 1970). Full backward induction is computationally very intensive, however. Carlin, Kadane, and Gelfand (1998) address stopping a fully sequential clinical trial from a Bayesian standpoint. They characterize the sequential design by a set of $2k + 1$ critical cutoffs for the posterior mean of a parameter that quantifies the advantage of a treatment over placebo. They present a forward simulation algorithm that substantially eases the analytic and computational burdens associated with backward induction. Brockwell and Kadane (2003) and Müller et al. (2007) construct a grid-based approximation to the expected loss at each decision time, viewing the expected loss at each decision time as a function of certain statistics of the posterior distribution of the parameter of interest.

We utilize a dual strategy of forward simulation and constrained backward induction, as in Müller et al. (2007), to maximize the expected utility and find the optimal sequential design. In principle, the dual strategy divides the continuous posterior mean $S_i = E(\theta_i | \text{data})$ into discrete units on a grid. The backward induction is constrained to the set of grid values. In general, any suitable low-dimensional summary statistic can be used—not just the mean. The statistic should capture most of the information in the full posterior distribution. Besides the constant sampling cost, the utility function is determined only by possible advantage of the new treatment over the standard one. This dependence on the success probabilities motivates considering the posterior mean of the success probability of the new treatment to be the summary statistic used in the constrained backward induction.

Recall that the number of decision time points within a given study has finite horizon T . One can simulate as many sample paths for a study as one chooses. We generate L possible experiments $w^l = (\theta^l, Y_t^l)$, $l = 1, \dots, L$, with $\theta^l \sim p(\theta)$, and $y^l \sim p(y | \theta^l)$ for all analysis-decision

time points $t = 1, \dots, T$. We consider replicates of a single study in the simulations, so we write generic θ without a subscript. For each simulated experiment w^l , we record the summary statistic S_t^l at $t = 1, \dots, T$.

We estimate the expected utility for each decision at each grid cell along the posterior mean axis by working backwards. For a grid cell at the time horizon T , we evaluate the expected utility for a possible action d_T as the sample average of the realized utilities under d_T . The average is over all forward simulations with a final value of the summary statistic falling within that grid cell at T . For $T-1$, we proceed similarly, except that now the possible actions include continuation. For any forward simulation with S_{T-1}^l falling within the grid cell under consideration, we evaluate the realized utility of continuation as the sampling cost for one more patient cohort plus the already computed expected utility under the optimal decision at time T . In this way, we work backwards and choose the decision that maximizes the expected utility at each grid cell for each decision-time point. We call this decision the optimal one.

The decision table consists of 1s, 2s, and 3s, as defined in Section 2.1. Figure 1 shows the decision boundaries for a design with analyses after cohorts of five patients and a maximum of forty patients per study (i.e., $n_1 = 5$ and $T = 8$). The X-axis is the analysis-decision time, and the Y-axis is the posterior mean of the success probability of the new treatment. The decision table indicates for each analysis-decision time point which decision is optimal, as a function of the posterior mean. If the posterior success probability of the new treatment is below the lower boundary, then it is optimal to stop the study and discontinue studying the new treatment; if the posterior mean is above the upper boundary, then the optimal decision is to stop the current study and switch to the phase III trial. Continuing the current study is optimal if the posterior mean is between the two boundaries at any analysis. As patient information accumulates, the continuation region becomes narrower, indicating greater certainty with more information. Please see the supplementary material for more information about how we generated this table.

At the last analysis-decision time we have to stop and choose between abandoning the drug and moving to phase III. In our example, it is better to move to a phase III trial if the posterior success probability is above 0.5 at the final analysis; if the posterior mean is less than 0.5, then the optimal decision is to abandon further development of the treatment.

An ideal decision table would have no islands within each region corresponding to a decision. That is, the region for decision 1 should not be contaminated with numbers 2 or 3, etc. Because of simulation and round-off errors, however, this is not always the case. We applied a linear spline at each analysis-decision time to smooth ragged decision boundaries.

For any particular maximum sample size per treatment, the decision table stays the same, as long as all the parameters in the utility function do not change. In particular, the decision table will not change if the prior parameters γ , σ^2 , and τ^2 in the hierarchical probit model (1) change. Changing the hyperparameters affects the posterior inference (e.g., posterior mean) but not the decision tables. Thus, the same decision table can be used for all treatments.

4. Evaluation of the Design

This section examines some simulated examples of the decision rules. We first compare our approach to Wang and Leung's fully sequential design. Then, we present simulation results over a range of values for the hyperparameters and discuss the sensitivity of the results to the selection of the prior distributions.

4.1 Comparison with Wang and Leung's Approach

We illustrate our method and compare it to the approach of Wang and Leung's for two reasons. Both methods take a decision-theoretic approach, and both are fully sequential. Making decisions after collecting data from a group of patients reduces the opportunity to stop the trial early. Compared to Yao, Begg, and Livingston, Wang and Leung's fully sequential design is about 50% more efficient in most cases. Efficiency here means that the design identifies the first promising treatment with fewer patients.

We carried out simulation studies to evaluate the frequentist operating characteristics of the design. We based our comparison on the expected number of patients treated until one identifies the first promising treatment and the probability that a treatment is confirmed to be efficacious after the decision to move it to a phase III trial.

In the simulations, we used a Beta distribution for the probability of success of the treatments under study. We used the same sets of parameters (α, β) as Wang and Leung to facilitate comparison with their results. For ease of comparison, we labeled the cases in Table 2 from Wang and Leung's paper from 1 to 16, consecutively by row. We also used the same target value of $\theta^* = 0.5$, i.e., we consider a treatment with at least 0.5 probability of success to be promising. Using results in Wang and Leung (1998), we determined the phase II sampling cost based on their minimum cost (their equation 2) divided by the maximum sample size. For example, the minimum cost in case 5 is 30.5. Dividing 30.5 by the maximum sample size in this case (30) gives $c_1 = 1.0167$. The phase III sampling cost was $c_2 = 5 \times c_1$, and we set the reward, b , empirically (roughly, $b = 1614 + 650 \times c_1$).

Wang and Leung (WL) assumed that the prior distribution of θ is uniform, i.e., $\text{Beta}(1,1)$. We chose the parameter values $\gamma = 0$, $\sigma^2 = 0.8^2$ and $\tau^2 = 0.8^2$ in the hierarchical probit model (1), so that the prior distribution of the probability of success of the first treatment looks like a uniform (0, 1) distribution (see Figure 2c).

We considered two decision errors similar to frequentist Type I and Type II error rates. We do not call these incorrect decisions Type I or Type II errors, since our two errors are measured using posterior distributions and are used in a Bayesian design. When we write "accept" or "accepted," we mean that the phase II study leads to the decision to go to a phase III trial. Similarly, "reject" or "rejected" means that the phase II study leads to the decision to discard the treatment. Intuitively, we would like a treatment to be called efficacious with high probability in the confirmatory phase III trial when the treatment is accepted in a phase II screening trial. We call the event that a treatment is deemed efficacious in phase III, given that the treatment was accepted in the phase II study, a "true positive" (TP). The probability of a TP is denoted PTP. We also want low probability of abandoning an efficacious treatment (i.e., rejecting). We call this probability PFN, i.e., the probability of a false negative decision in the phase II study. We summarize these terms in Table 1.

One might wish to incorporate PFP and PFN in the utility function, although we do not in our example. FP errors may result in further patients being treated with an ineffective treatment, while FN errors may have the effect of discouraging further experiments with a truly effective treatment. From a drug development perspective, FN is a more serious mistake, because it may lead one to overlook an effective treatment (see Simon, 1989).

Table 2 summarizes the simulation results comparing our method to WL. We found that our method and WL had roughly the same expected number of treatments to study until finding the first promising treatment. Our proposal, however, required a smaller number of patients, on average, before identifying the first promising treatment (N_p), except in cases 2, 4, and

13. The PTPs are quite similar for the method of Wang and Leung and our method (Table 2). PFNs are close in cases with $E(\theta)$ equal to 0.2 and 0.3 (the third and fourth columns in Table 2). In cases 4, 8, 12 and 16, (i.e., Beta models for θ with mean 0.5), however, our design provided smaller PFNs. Recall that we use $\theta^* = 0.5$ as the target value in the phase III confirmatory trial. It seems that when the mean probability of success for θ is close to the target value, the method we propose has better design characteristics.

4.2 Sensitivity of Posterior Inference to the Prior

We investigated the sensitivity of the design's operating characteristics to changes in the prior distributions of the hyperparameters γ , σ^2 , and τ^2 in model (1). We considered the case of a fully-sequential design with maximum time horizon of 30 patients per study (or per treatment). In the simulations, a treatment's success probability followed a Beta (0.12, 0.48) distribution (mean = 0.2 and variance = 0.1). This distribution, used for illustration, assigns probability 0.71 to values less than or equal to 0.2 and probability 0.18 to values greater than or equal to 0.5.

We evaluated the optimal design across several ranges of values for the hyperparameters. Figure 2 displays prior distributions of the success probabilities for the cases in which γ equals $\{-0.5, 0, 0.5\}$ and σ^2 and τ^2 each equal $\{0.3^2, 0.5^2, 0.8^2\}$. In each subplot, as γ increases from -0.5 to 0 to 0.5 , the prior density of the success probability shift from the left to the right, meaning that the mean probability of success increases. Looking across subplots in Figure 2, as the variance terms σ^2 and τ^2 increase, the prior densities of the success probability become flatter, meaning that the density has less precision around the prior mean probability of success. A uniform (0,1) prior corresponds roughly to the case of $\gamma = 0.0$ and $\sigma^2 = \tau^2 = 0.8^2$.

Table 3 shows that the expected number of patients treated before one identifies a promising agent increases as γ increases. In the cases where σ^2 and τ^2 equal 0.3^2 , the expected number of patients to treat before identifying the first promising treatment increases from 31.25 to 48.15 to 54.05 as γ increases from -0.5 to 0 to 0.5 . The expected number of treatments decreases, however, as do PTP and PFN.

Recall that the underlying probability of success of the treatments to test has mean 0.2 in the simulations. If we start with an optimistic prior, i.e., a prior with mean probability of success higher than the mean probability of success of the treatments to be tested, it takes longer to identify treatments with low efficacy. Because it takes longer, the expected number of patients required increases, but the expected number of treatments decreases slightly. Also, with an optimistic prior, there is a slightly smaller chance of falsely rejecting an effective treatment. This lower chance of falsely discarding a good treatment comes at the cost of a higher risk of picking out treatments that turn out not to be efficacious.

Table 3 shows that for a value of γ , as the variance terms σ^2 and τ^2 increase, the expected number of patients decreases. The higher the precision around the prior mean is, the larger the expected number of patients required. In the case of $\gamma = -0.5$, the prior expected value of θ_1 ranges from 0.35 ($\sigma^2 = \tau^2 = 0.8^2$) to 0.31 ($\sigma^2 = \tau^2 = 0.3^2$). If, *a priori*, the treatments have low mean probability of success, we need more patients to identify a treatment with efficacy, increasing the expected number of patients to enroll before finding the first promising treatment. In the case $\gamma = 0$, the prior mean probabilities of success are very close to the target value 0.5. It takes longer to detect both inefficacious and promising treatments. When $\gamma = 0.5$, the prior mean probability of success is larger than the target value 0.5. This case shows that it will take longer to reject treatments with low efficacy (and increase the expected sample size) if the initial prior has relatively high precision and is centered at a probability of success greater than the target value. PFNs are larger in the case of $\gamma = -0.5$

than those of $\gamma = 0$ and 0.5. We infer from our example that if the treatments have low mean success probabilities and the initial prior has mean probability of success closer to that of the treatments, then there is a greater chance of falsely rejecting an efficacious one. We may reject slightly more treatments than we desired.

5. Discussion

We applied Bayesian decision theory to phase II screening trials that consist of a sequence of studies. Each study may concern a new drug, a new combination of drugs, or a new indication for an existing drug, for example. With our approach, one can produce a decision table at the start of the sequence of studies for use throughout. The methodology employed forward simulation of study data, a low-level summary statistic to characterize uncertainty about the overall parameter space, and backward induction over a grid. A utility function characterized appropriate costs and benefits. The optimal decision at any time depended only on the value of a low-level summary statistic (or statistics) at that time, no matter how one got there. One can also incorporate covariates via the hierarchical probit model. If two or more studies happen to evaluate the same treatment, then the hierarchical model would allow linking the studies.

Bayesian decision-making, although computationally intensive, provides an efficient and rational approach to the phase II screening process. We want to find active treatments and rule out clearly ineffective therapies but do not want to rush to a premature decision based on a small sample. Hierarchical modeling allows borrowing strength across studies and improves precision of estimation. The resulting shrinkage is advantageous, because the decisions to abandon a treatment or to switch to pivotal trial are more conservative. Being conservative in this setting is important, since we typically know little about the true nature of the treatments.

Through the Bayesian hierarchical model, we continually update our knowledge about the underlying population distribution of treatment-specific success probabilities. A consequence of updating is that the expected number of patients depends on the data observed in those studies. The order in which the potential treatments arrive also plays an important role. Because treatment order matters, it might be sensible to reconsider previously rejected treatments at later stages. An optimal ordering, on the other hand, would choose the next treatment for evaluation so as to maximize the expected utility from the start of the whole process. The computation would become more complicated, however. This problem may be solvable for simple utility functions, but the backward induction is generally formidable. A myopic strategy that would consider at most r future treatments could provide an approximate solution.

We pre-specify the maximum sample size per trial to enable the backward induction. Typically, the maximum sample size is based on an assumed patient accrual rate, monetary costs, and the statistical properties of the design. A typical phase II clinical trial requires 30 to 80 patients. Trials with more than 100 patients are often impractical in terms of timing and cost of the trials. When using our method, one can consider frequentist properties when choosing the maximum sample size per study.

We presented the method in the context of a sequence of studies, one after another, for ease of exposition. We mainly envision the use of the method in this sequential setting. One could easily expand the approach to allow for contemporaneous studies, with $i \in I(t)$, the set of studies at time t . Still, we envision making decisions in each study separately, with decisions not directly reflecting head-to-head comparisons among the ongoing studies. One

might also consider an evolving process in which the target response rate changes over time. We are currently investigating this enhancement.

We set the potential reward (b) and phase II and phase III sampling costs somewhat arbitrarily. Although these numbers would be available in practice, we can provide some idea of the effect of these parameters on the decision table. In terms of the sampling costs, what matters are the products $n_1 * c_1$ and $n_2 * c_2$. The phase III sample size (n_2) is set according to the desired power. We focus our discussion, therefore, on changing c_2 (cost per patient in the phase III study) relative to c_1 (cost per patient in the phase II study).

If the phase III cost per patient increases relative to the per-patient cost in phase II, then it is relatively cheaper to continue than to move to phase III. The slope of the upper boundary of the decision table (Figure 1) will get steeper (i.e., more negative). For points near the current lower boundary (between “Continue” and “Discard”), the probability of the possible benefit in phase III will decline relative to the cost, so it will be better to stop rather than continue. The lower boundary of the wedge formed by the stopping boundaries will move up. Thus, as c_2 increases relative to c_1 , the wedge will move up and rotate clockwise.

Similarly, as c_2 gets smaller relative to c_1 , we observe the opposite effect on the boundaries. The cost of the phase III study gets cheaper, so there is less to lose by taking the risk. Now, the upper boundary (“go to phase III”) moves down, and the wedge rotates counter-clockwise.

Decisions also depend on the value of the reward (b). If the reward is too low, then there is little point in even carrying out the phase II study. As the reward gets greater, relative to the sampling cost in phase II, the boundary between continuing the phase II study and moving to phase III will move down, in a manner similar to reducing c_2 relative to c_1 .

We set the hypervariances (σ^2 and τ^2) equal to each other. In the absence of more information, we would recommend starting with the roughly uniform hyperprior ($\gamma = 0.0$ and $\sigma^2 = \tau^2 = 0.8^2$). If one makes the variances too large (e.g., ≥ 2), then one may encounter numerical problems. Setting $\sigma^2 = \tau^2$ does not appear to restrict the set of possible priors, and one can find several combinations that will yield similar results. Naturally, relevant prior information characterizing study-to-study variation should influence the value of σ^2 .

We envision an iterative implementation of the design. That is, one would set key parameters to some reasonable values, consider the performance of the resulting design via summaries (as discussed in Section 4), and revise the parameters as appropriate.

In the discussion above, we used drug development as the paradigm. The proposed methodology has broader application, however. It is applicable in any setting in which one wishes to screen several contending innovations that appear over time. The utility functions will change to reflect the circumstances.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank an Associate Editor for many helpful suggestions. This research was partially supported by grants CA97534 and CA75981 from the U.S. National Cancer Institute.

References

- Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 1993; 88:669–679.
- Berry DA, Pearson LM. Optimal designs for clinical trials with dichotomous response. *Statistics in Medicine*. 1985; 4:497–508. [PubMed: 4089351]
- Brockwell A, Kadane J. A gridding method for Bayesian sequential decision problems. *Journal of Computational & Graphical Statistics*. 2003; 12:566–584.
- Carlin B, Kadane J, Gelfand A. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*. 1998; 54:964–975. [PubMed: 9750245]
- DeGroot, MH. *Optimal Statistical Decisions*. McGraw-Hill Inc.; New York: 1970.
- Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*. 1961; 13:346–353. [PubMed: 13704181]
- Heitjan DF. Bayesian Interim Analysis of Phase II Cancer Clinical Trials. *Statistics in Medicine*. 1997; 16:1791–1802. [PubMed: 9280033]
- Leung DHY, Wang YG. Optimal designs for evaluating a series of treatments. *Biometrics*. 2001; 57:168–171. [PubMed: 11252593]
- Müller P, Berry D, Grieve A, Smith M, Krams M. Simulation-Based Sequential Bayesian Design. *Journal of Statistical Planning and Inference*. 2007 (in press).
- Rossell D, Müller P, Rosner GL. Screening Designs for Drug Development. *Biostatistics*. 2007; 8:595–608. [PubMed: 17030589]
- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
- Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*. 1999; 55:971–977. [PubMed: 11315037]
- Stout QF, Hardwick J. Optimal screening designs with flexible cost and constraint structures. *Journal of Statistical Planning and Inference*. 2005; 132:149–162.
- Sylvester RJ. A Bayesian approach to the design of phase II clinical trials. *Biometrics*. 1988; 44:823–836. [PubMed: 3203131]
- Thall P, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*. 1994; 50:337–349. [PubMed: 7980801]
- Wang YG, Leung DHY. An optimal design for screening trials. *Biometrics*. 1998; 54:243–250. [PubMed: 9544519]
- Yao TJ, Begg CB, Livingston PO. Optimal sample size for a series of pilot trials of new agents. *Biometrics*. 1996; 52:992–1001. [PubMed: 8805764]
- Yao TJ, Venkatraman ES. Optimal two-stage design for a series of pilot trials of new agents. *Biometrics*. 1998; 54:1183–1189. [PubMed: 9750258]

Figure 1. shows the decision boundaries for a group sequential design with cohort size 5 and maximum 40 patients per study. The parameters in the model were $\gamma = 0.0$; $\sigma^2 = \tau^2 = 0.8^2$; $n_1 = 5$; $T = 8$; $c_1 = 0.14$ units; $c_2 = 0.7$ units; $b = 290$ units; $n_2 = 96$.

Figure 2. Prior distributions of the success probability of the first treatment with different γ and the same $\sigma = \tau$ in each subplot. The three parameters in the legend are γ , σ , and τ , respectively.

Table 1

2×2 table for Bayesian errors

		Phase III (Efficacious?)		
		Yes	No	Total
Phase II	Accept	PTP	PFP	1
	Reject	PFN	PTN	1

PTP = $P(\text{Phase III Yes} \mid \text{Phase II Accepted})$ PFP = $1 - \text{PTP}$

PFN = $P(\text{Phase III Yes} \mid \text{Phase II Rejected})$ PTN = $1 - \text{PFN}$

Table 2

Comparison of the method of Wang and Leung (WL) and our proposed Bayesian optimal design approach (OPT). Both method used the same maximum sample size (T).

Var (θ)		E (θ)			
		0.2	0.3	0.4	0.5
0.08		1) ^a	2)	3)	4)
	$E_{WL}[N_p]^b$	32	35	28	16
	$E_{OPT}[N_p]^c$	27	40	28	17
	Max T^d	60	130	100	60
	PTP_{WL}, PTP_{OPT}	0.96, 0.95	0.93, 0.92	0.96, 0.94	0.96, 0.94
	PFN_{WL}, PFN_{OPT}	0.08, 0.06	0.13, 0.11	0.22, 0.18	0.28, 0.18
0.10		5)	6)	7)	8)
	$E_{WL}[N_p]$	17	22	17	11
	$E_{OPT}[N_p]$	13	15	14	11
	Max T	30	40	40	60
	PTP_{WL}, PTP_{OPT}	0.93, 0.95	0.95, 0.94	0.95, 0.93	0.95, 0.97
	PFN_{WL}, PFN_{OPT}	0.07, 0.06	0.12, 0.10	0.21, 0.17	0.26, 0.16
0.12		9)	10)	11)	12)
	$E_{WL}[N_p]$	12	13	11	8
	$E_{OPT}[N_p]$	6	10	10	7
	Max T	30	40	20	30
	PTP_{WL}, PTP_{OPT}	0.95, 0.96	0.96, 0.98	0.94, 0.98	0.96, 0.98
	PFN_{WL}, PFN_{OPT}	0.04, 0.03	0.10, 0.08	0.17, 0.15	0.24, 0.19
0.14		13)	14)	15)	16)
	$E_{WL}[N_p]$	5	9	7	6
	$E_{OPT}[N_p]$	8	7	6	5
	Max T	5	10	10	10
	PTP_{WL}, PTP_{OPT}	0.95, 0.98	0.92, 0.93	0.94, 0.97	0.95, 0.96

Var (θ)	E (θ)				
	0.2	0.3	0.4	0.5	
PFN_{WL}, PFN_{OPT}	0.03, 0.02	0.08, 0.10	0.12, 0.11	0.18, 0.16	

- ^aCase number (see text).
- ^bExpected number of patients until identify promising treatment (N_P) for WL (Wang and Leung, 1998).
- ^cExpected number of patients until identify promising treatment (N_P) for our design.
- ^dMaximum sample size per study (T) from a table in Wang and Leung (1998).
- ^eProbability of a True Positive (PTP) decision for WL and OPT.
- ^fProbability of a False Negative (PFN) decision for WL and OPT.

Comparison of the expected number of patients until identifying a promising treatment (N_p), the expected number of treatments until identification of the first active treatment, PTP, and PFN for priors with different combinations of γ , σ^2 , and τ^2 . Simulated treatment-specific success probabilities followed a Beta(0.12, 0.48) distribution.

Table 3

γ	σ^2 & τ^2	$Mean(\theta_i)$	$\Pr(\theta_i \geq 0.5 \gamma, \tau^2, \sigma^2)$	$\Pr(\theta_i \leq 0.2 \gamma, \tau^2, \sigma^2)$	N_p	No. of Treatments	PTP	PFN
-0.5	0.3 ²	0.32	0.112	0.208	31.25	5.93	0.98	0.08
	0.5 ²	0.34	0.241	0.307	21.53	6.21	0.98	0.07
	0.8 ²	0.37	0.328	0.386	10.37	6.01	0.96	0.08
	0.3 ²	0.50	0.503	0.025	48.15	5.57	0.98	0.05
0	0.5 ²	0.50	0.502	0.119	27.38	5.39	0.98	0.05
	0.8 ²	0.50	0.504	0.227	12.76	5.28	0.95	0.06
	0.3 ²	0.68	0.886	0.001	54.05	4.91	0.97	0.03
	0.5 ²	0.66	0.762	0.029	31.04	5.01	0.96	0.04
0.5	0.8 ²	0.63	0.672	0.117	19.84	4.95	0.94	0.05