Bayesian Optimal Design for Phase II Screening Trials

Meichun Ding [1,*], Gary L. Rosner [2], and Peter Müller [2]

[1] Department of Biostatistics and Epidemiology, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA, 91320

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

[*] Email: mding@rice.edu

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. 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

Many authors have proposed phase II clinical study designs in the literature over the last three decades, both frequentist and Bayesian approaches. 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 designs that fix the total sample size. 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. The sample size in each study is fixed, but the total sample size in the 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 in the same setup. Taking a Bayesian decision-theoretic approach, Rossell, Müller, and Rosner (2006) find optimal linear boundaries for fully sequential phase II screening studies. Stout and Hardwick (2005) present a general decision-theoretic framework for finding optimal designs for screening trials.

A common feature of these designs is the assumption 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 statistical theory and driven by the desire to reduce the risk of making incorrect decisions with respect to statistical hypothesis tests: either deciding in favor of H_0 when H_1 is the true state of nature or deciding in favor of H_1 when H_0 is correct. In other words, these proposals seek to minimize frequentist error probabilities.

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., Sylvester and Staquet (1977,

1980), Sylvester (1988), Berry and Pearson (1985), Brunier and Whitehead (1994), Berry and Ho (1988), Stallard (1998), Stallard, Thall, and Whitehead (2000), and Stallard and Thall (2001). These designs are based on maximizing the expected value of a pre-specified utility function. The second Bayesian approach does not explicitly specify a loss function or cost of sampling. Instead, stopping criteria are based on whether the posterior probability of some clinically important event exceeding 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.

Whether one considers a frequentist or Bayesian approach, most screening design proposals in the oncology literature so far consider one treatment at a time. Each study is considered 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 the design of 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. There is a sampling cost per patient incurred in collecting the data, as well as 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. 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 where 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 is with respect to 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. We assume that the response is observable soon enough to enable the sequential designs.

We define the algorithm for a screening process that considers multiple studies simultaneously. We present the method, however, in the context of a sequence of studies, one after another, for ease of exposition. Besides, we mainly envision the use of the method in this sequential setting.

At each analysis-decision time t, we choose a decision d_{ti} for the current study. Assume $d_{ti} \in \{1,2,3\}$. Decision $d_{ti} = 1$ indicates abandoning the treatment in study i ("stop and discard the experimental treatment"); $d_{ti} = 2$ indicates that the current study (i) should stop and the treatment should progress to a phase III trial ("stop and switch to pivotal trial"); decision $d_{ti} = 3$ indicates continuation with new patients entering the study in period $t+1$ ("continue sampling on the same experimental treatment"). Let $d_t = (d_{t1},\ldots,d_{tk})$, with the understanding that $d_t = d_{t1}$ when considering only one study at a time, i.e., k=1. Let $D_t = (d_1,...,d_t)$ be the decision on all the treatments up to time t. We let $Y_t = (y_1,...,y_t)$ be all the data observed for all the patients and treatments up to time t, where $y_t = (y_{tij}; i = 1,...,k; j = 1,...,n_{ti})$. Define $H_t = (D_{t-1}, Y_t)$ as all the decisions up to time $t-1$ and all the data up to time t. We call H_t the history up to time t. Decisions d_t can depend on all the data up to time t and all the decisions up to time $t-1$, i.e., $d_t = d_t(H_t)$.

Each study is a single-arm design in that we do not compare treatments. The choice of a single-arm design has been made for two reasons. First, classical two-arm trials often require larger sample sizes than are practical for phase II studies. Second, the goal of many phase II trials, as 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 and its 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 observed data for patient j on treatment i at analysis-decision t is binomially distributed, $y_{tij} \mid \theta_i \sim Bin(1,\theta_i)$. 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 $\Phi(t)$ denotes the standard normal cumulative distribution function. The next level of the hierarchy characterizes 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 μ_0. At the bottom level, uncertainty about the hyper-parameter μ_0 is also characterized by a normal distribution. The hierarchical probability model can be written as:

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

The first equation in the hierarchical probit model represents the likelihood. The last three equations represent the prior distribution of the treatment- or 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 hyper-prior normal distribution with parameter μ_0, we continually update our knowledge about μ_0 over time

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 transformation that is more amenable to the implementation of stochastic posterior simulation, as in Albert and Chib (1993). We introduce latent variables z_{tij} for each patient j on treatment i at the analysis-decision time point t. In mathematical form, we can write $p(z_{tij} \mid \mu_i) = N(\mu_i, 1)$ with

$y_{tij} = \{1 \text{ if } z_{tij} \geq 0;\ 0 \text{ otherwise}\}$. Since

$P(y_{tij} = 1) = P(z_{tij} \geq 0) = P(\frac{z_{tij} - \mu_i}{1} \geq -\mu_i) = 1 - \Phi(-\mu_i) = \Phi(\mu_i)$, we get $y_{tij} \sim Bin(1, \theta_i)$, as desired.

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 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. We define b as the reward for one unit improvement of the success probability of the new treatment over the old or current standard treatment. We define random variable $\overline{\Delta}$ as the difference of success probabilities (new treatment minus the old one, i.e., $\theta_{new} - \theta_{old}$). If we decide to initiate a pivotal phase III trial and this trial concludes that the drug is, in fact, an effective treatment, there is a positive benefit $b * \overline{\Delta}$. 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 $\max(0, b * \overline{\Delta}) \times I_{\{\text{significant outcome}\}}$, with $I_{\{A\}}$ and indicator function for event A. The phase III benefit in the utility calculation is a probabilistic calculation involving the 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 all the treatments under active evaluations in phase II. 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. Typical requirements: at most a 5% chance of wrongly deciding that the new treatment is better (Type I error rate α) and at least a 90% chance of detecting a clinically relevant difference $\delta > 0$ when it is present (power of the test, $1 - \beta$). We set the difference δ to 0.3 in our example. The success probability of the standard or control treatment θ_{old} is set to 0.2 for the purpose of illustration. In the future phase III trial, we randomize n_2 patients equally to the new treatment and the standard one, testing $H_0 : \theta_{new} = \theta_{old}$ vs. $H_a : \theta_{new} > \theta_{old}$. 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.

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 distribution having mean 0.2 and variance 0.002, i.e., beta (20, 80). 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 to be too large.

The test statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{\frac{n_2}{2}} + \frac{\hat{p}_2(1-\hat{p}_2)}{\frac{n_2}{2}}}},$$

where $\hat{p}_1$ and $\hat{p}_2$ are observed proportions of successes for the new treatment and the standard one, respectively. These statistics are the usual frequentist estimates of the treatment-specific success probabilities θ_{new} and θ_{old} after conducting the phase III study. The standard one-sided test calls the new treatment efficacious if $z > z_{1-\alpha}$.

The utility function can be expressed as follows.

$$u_t(d_t, \theta, Y_t, Y_{III}) = \begin{cases} -c_1 * n_1 * t & \text{if stop to discard} \\ -\{c_1 * n_1 * t + c_2 * n_2\} + & \\ \quad b * \{\theta_{new} - \theta_{old}\} I_{[z > z_{1-\alpha}]} & \text{if stop to switch to phase III trial} \end{cases} \quad (2)$$

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 sampling algorithm that substantially eases the analytic and computational burdens associated with backward induction. Brockwell and Kadane (2001) and Müller et al. (2006) 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 proposed by Berry et al. (2001), to maximize the expected utility and find the optimal sequential design. In principle, the dual strategy divides the continuous posterior mean into discrete units. The backward induction is constrained to a set of values on a grid. We use a grid on the posterior mean $S_i = E(\theta_i | \text{data})$. In general, any other suitable low dimensional summary statistic could be used. The statistic should capture most of the information in the full summary statistics. Besides the constant sampling cost, the utility function is determined only by possible advantage of the new treatment over the standard one. This motivates considering the posterior mean of the new treatment to be the summary statistic S_t used in the constrained backward induction.

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

$t = 1,...,T$. For each simulated experiment w^i, we record the summary statistic S_t^i at $t = 1,...,T$. Starting with the last analysis-decision time point and then working backwards, we choose the decision that maximizes the expected utility at each analysis-decision time point. We call this decision the optimal one.

We estimate the expected utility for each decision (1, 2, and 3) at each grid cell along the posterior mean axis by working backwards. The decision table consists of 1s, 2s, and 3s, with 1 indicating the decision "stop and discard the experimental treatment", 2 indicating "stop and switch to phase III," and 3 indicating "continue sampling on the same 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, however, this is not always the case. We utilized a linear spline to smooth ragged decision boundaries within each time to solve this problem.

Figure 1 shows the decision boundaries for a design with analyses after cohorts of five patients and a maximum of forty patients per study. 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, depending on 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 an analysis. As patient information accumulates, the continuation region becomes narrower, indicating greater certainty with more information.

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

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 hyper-parameters only affects the posterior inference, 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 with a noninformative prior 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 a small number of patients are needed to identify the first promising treatment.

We base 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 with parameters a and b for the probability of success of the treatments under study. We use the same combination of (a, b) as did Wang and Leung to facilitate comparison with their results. We also use 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.

Wang and Leung (WL) assume that the prior distribution of θ is uniform, i.e., beta(1,1). We choose 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 (cf. Figure 1). We label the cases in Table 2 from Wang and Leung's paper as case 1 to 16, consecutively by row, for ease of comparison. We carry out simulation studies to evaluate the frequentist operating characteristics of the design.

We consider two concepts similar to frequentist Type I and Type II error rates. We call these decision errors, rather than Type I or Type II errors, since our two errors are measured using posterior distributions and are used in a Bayesian design. We first define the terms "accept" and "reject" to avoid potential confusion with the same terms in hypothesis testing. 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 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) in the phase II screening trial. 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 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 screen out the first promising treatment. Our proposal, however, required a smaller number of patients, on average, to screen out the first promising treatment, except in cases 2, 4, and 13. We saw similar results when we examined the expected number of treatments until one finds a promising one (not shown). The PTPs are quite similar for the method of Wang and Leung and our method (Table 3). PFNs are close in cases with $E(\theta)$ equal to 0.2 and 0.3 (the first two columns in Table 4). In cases 4, 8, 12 and 16, (i.e., Beta models for θ with mean 0.5), however, our design provided noticeably 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 hyper-parameters 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). For the purpose of illustration, we used a Beta(0.12, 0.48) distribution (mean = 0.2 and variance = 0.1) when generating a treatment's success probability in the simulations. This beta distribution 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 density plots of prior distributions of success probability become flatter, meaning that the density has less precision around the prior mean probability of success. A uniform prior corresponds roughly to the case of $\gamma = 0.5$ 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 identify 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 screen out 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 the same γ value, 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, it takes time to screen out the treatment with efficacy, increasing the expected number of patients to screen out the first promising treatment. In the case of $\gamma = 0$, the prior mean probabilities of success are very close to the target value 0.5. It takes longer to screen out both inefficacious and promising treatments. In the case of $\gamma = 0.5$, 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 we start with a prior having relatively high precision and 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. From our example, we see that if the treatments have low mean success probabilities and we start with a prior that has mean probability of success closer to that of the treatments, then there

is a greater chance of falsely rejecting an efficacious one. In this case, we may reject slightly more treatments than we might have wanted.

5. Discussion

In this paper we have applied Bayesian decision theory to phase II screening trials. Our approach led to a decision table one can produce at the start of the sequence of trials and 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, incorporating a utility function that characterizes appropriate costs and benefits. Because of the use of the low-level summary statistic (or statistics), the optimal decision at any time depends only on the value of this statistic at that time, no matter how one got there. Our method allows one to incorporate covariates via a hierarchical probit model.

Bayesian decision-making, although computationally intensive, provides an efficient and rational approach to the phase II screening process. We want to rule out clearly ineffective treatments. We do not, however, want to rush to a premature decision that a treatment is promising or useless based on a small sample. Bayesian hierarchical models allow us to borrow strength from other treatments and improve estimation in terms of better precision. The success probabilities are shrunk toward the overall average. The shrinkage is advantageous because the decisions to abandon a treatment or to switch to pivotal trial are more conservative. Being conservative in phase II screening trials is important when we are not sure about the true nature of the treatments and what works or does not work.

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 to be evaluated arrive also plays an important role. An optimal ordering would be one in which the next treatment for evaluation is always chosen so as to maximize the expected utility from the start of the whole process. The computation becomes more complicated, however. This problem may be solvable for simple utility functions, but the backward induction is generally formidable. A myopic strategy in which at most r future treatments would be considered can provide an approximate solution. Because ordering matters, it might be sensible to reconsider previously rejected treatments at later stages.

In order to enable the backward induction, we need to pre-specify the maximum sample size per trial. In practice, the maximum sample size is determined based on assumed patient accrual rate, feasible trial duration, and monetary costs, in addition to the statistical properties of the design and reliability of parameter estimates. 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. We can consider frequentist properties when choosing a maximum sample size.

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

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.

References

Albert, J.H and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.

Berry, D. A. and Ho, C.-H. (1988). One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* **44**, 219-227.

Berry, D. A. and Pearson, L. M. (1985). Optimal designs for clinical trials with dichotomous response. *Statistics in Medicine* **4**, 497-508.

Berry, D. A., Müller, P., Grieve, A. P., Smith, M., Parke, T., Blazek, R., Mitchard, N. & Krams, M. (2001). Adaptive Bayesian designs for dose-ranging drug trials. In C. Gatsonis, R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli & M. West, eds., *Case Studies in Bayesian Statistics, Volume V*, Lecture Notes in Statistics. New York: Springer-Verlag.

Brockwell, A. and Kadane J. (2003). A gridding method for Bayesian sequential decision problems. *Journal of Computational & Graphical Statistics*, **12**, 566-584.

Brunier, H. C. and Whitehead, J. (1994). Sample size for phase II clinical trials derived from Bayesian decision theory. *Statistics in Medicine* **13**, 2493-2502.

Carlin, B., Kadane, J., and Gelfand, A. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* **54**, 964-975.

DeGroot, M. H. (1970). *Optimal statistical decisisons*. McGraw-Hill Inc., New York.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J. Chron. Dis.* **13**, 346-353.

Heitjan, D. F. (1997). Bayesian Interim Analysis of Phase II Cancer Clinical Trials. *Statistics in Medicine* **16**, 1791-1802.

Leung, D. H. Y. and Wang, Y. G. (2001). Optimal designs for evaluating a series of treatments. *Biometrics* **57**, 168-171.

Müller, P., Berry, D., Grieve, A., Smith, M., and Krams, M. (2006). Simulation-Based Sequential Bayesian Design. Journal of Statistical Planning and Inference, (to appear).

Rossell, D., Müller, P., and Rosner, G.L. (2006). Screening Designs for Drug Development. Technical Report

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* **54**, 279-294.

Stallard N, Thall PF, Whitehead J (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **55**, 971-977.

Stallard, N. and Thall, P. F. (2001). Decision-theoretic designs for pre-phase II screening trials in oncology. *Biometrics* **57**, 1089-1095.

Stout, Q.F. and Hardwick, J. (2005). Optimal screening designs with flexible cost and constraint structures. *Journal of Statistical Planning and Inference* **132**, 149-162.

Sylvester, R. J. (1988) A Bayesian approach to the design of phase II clinical trials. *Biometrics* **44**, 823-836.

Sylvester, R. J. and Staquet, M. J. (1977). An application of decision theory to phase II clinical trials in cancer. *Recent Advances in Cancer Treatments* **35**, 342-356.

Sylvester, R. J. and Staquet, M. J. (1980). Design of phase II clinical trials in cancer using decision theory. *Cancer Treat. Rep.* **64**, 519-524.

Thall, P. and Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* **50**, 337-349.

Wang, Y. G. and Leung, D. H. Y. (1998). An optimal design for screening trials. *Biometrics* **54**, 243-250.

Yao, T. J., Begg, C. B., and Livingston, P. O. (1996). Optimal sample size for a series of pilot trials of new agents. *Biometrics* **52**, 992-1001.

Yao, T. J., and Venkatraman, E. S. (1998). Optimal two-stage design for a series of pilot trials of new agents. *Biometrics* **54**, 1183-1189.

Figure 1 shows the decision boundaries for a group sequential design with cohort size 5 and maximum 40 patients per study.

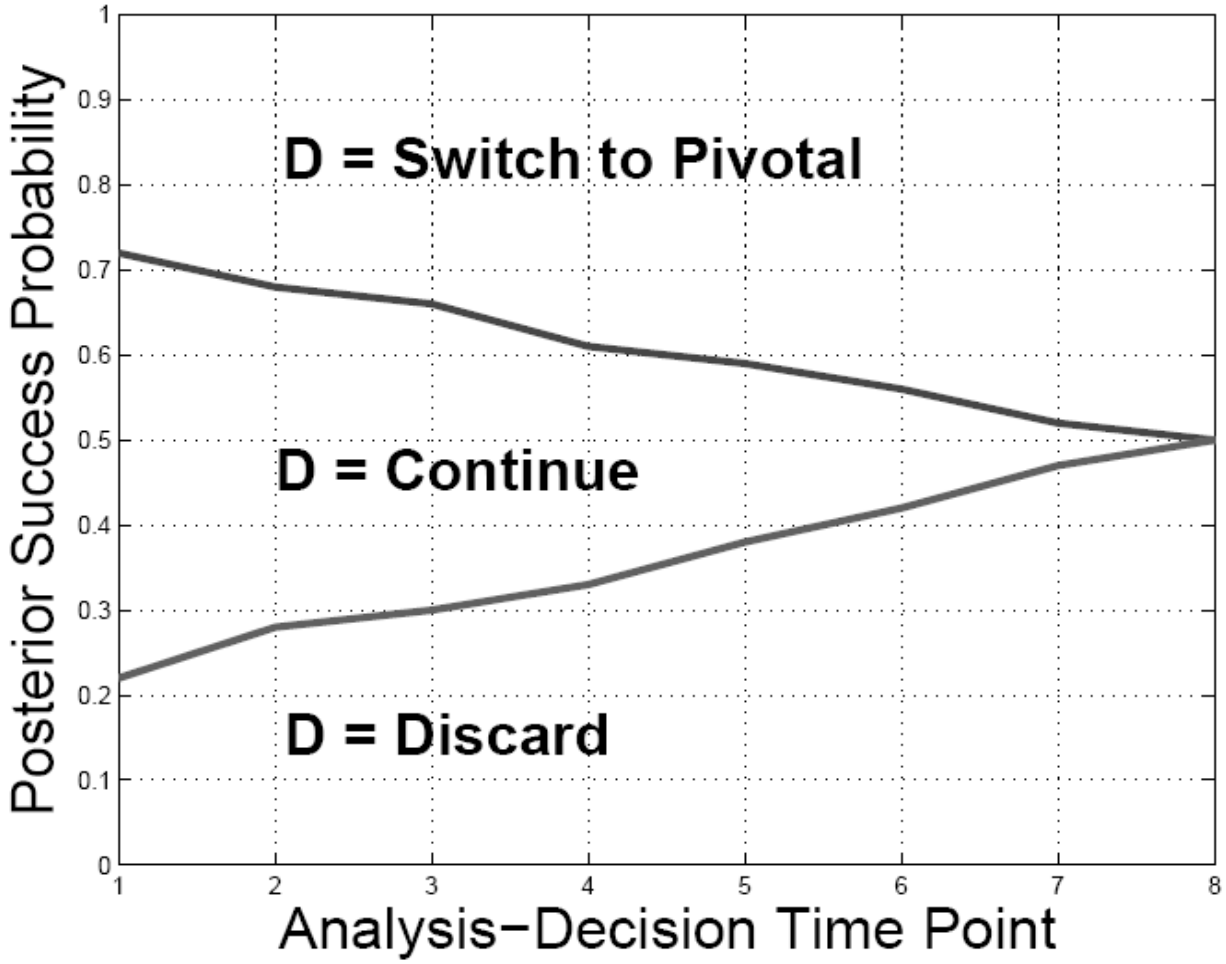

Figure 2. Prior distributions of the success probability of the first treatment with different γ and the same σ and τ in each subplot. The three parameters in the legend are $\gamma, \sigma,$ and τ, respectively.

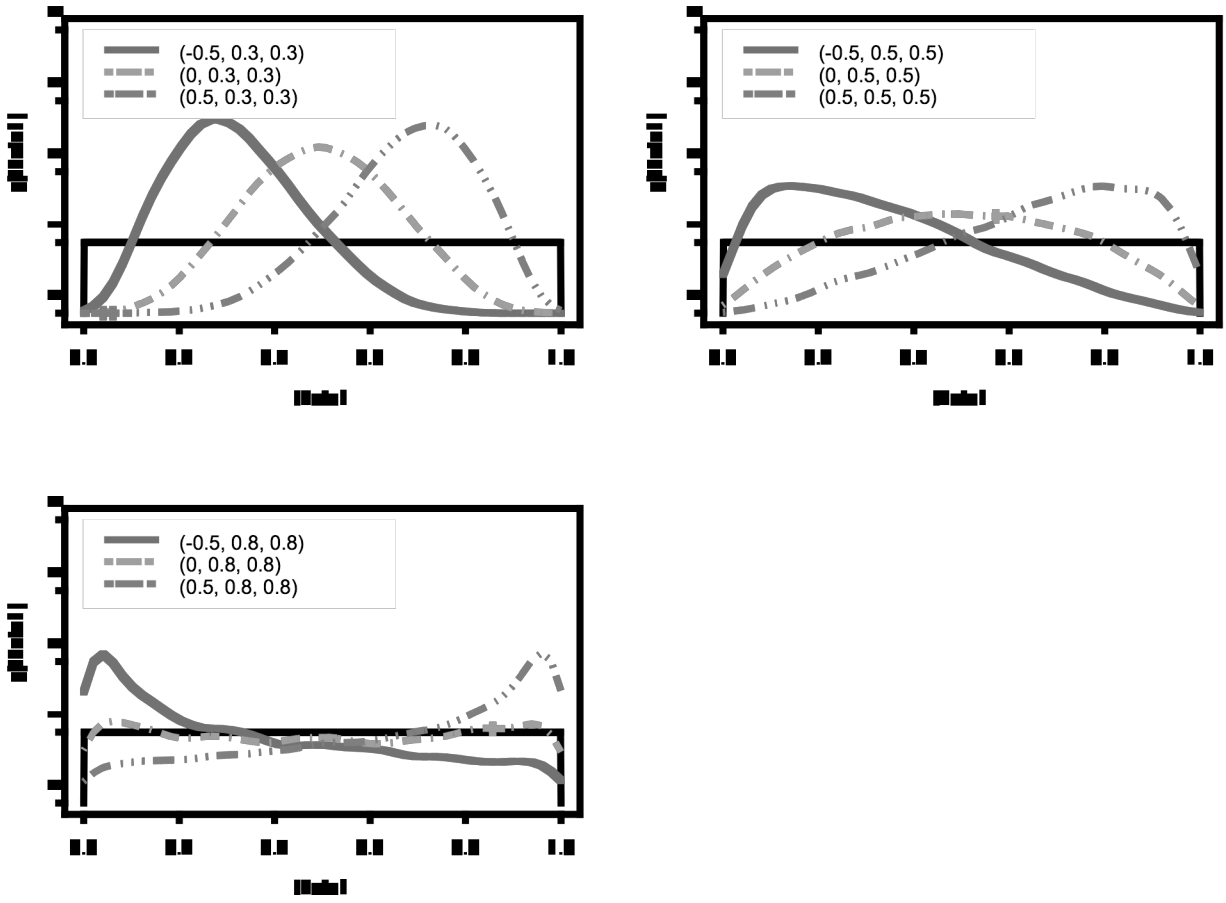

Table 1: 2 x 2 table for Bayesian errors

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

PTP = P (Phase III Yes | Phase II Accepted) PFP = 1-PTP

PFN = P (Phase III Yes | Phase II Rejected) PTN = 1-PFN

Table 2: Comparison of the method of Wang and Leung (WL) and our proposed Bayesian optimal design approach (OPT).

Var (θ)		E (θ) 0.2	0.3	0.4	0.5
		1)[a]	2)	3)	4)
0.08	$E_{WL}[N_p]$[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
		5)	6)	7)	8)
0.10	$E_{WL}[N_p]$	17	22	17	11
	$E_{OPT}[N_p]$	13	15	14	11
	Max T[d]	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
		9)	10)	11)	12)
0.12	$E_{WL}[N_p]$	12	13	11	8
	$E_{OPT}[N_p]$	6	10	10	7
	Max T[d]	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
		13)	14)	15)	16)
0.14	$E_{WL}[N_p]$	5	9	7	6
	$E_{OPT}[N_p]$	8	7	6	5
	Max T[d]	5	10	10	10
	PTP_{WL}, PTP_{OPT}	0.95, 0.98	0.92, 0.93	0.94, 0.97	0.95, 0.96
	PFN_{WL}, PFN_{OPT}	0.03, 0.02	0.08, 0.10	0.12, 0.11	0.18, 0.16

[a] Case number (see text).

[b] Expected N_p for WL (Wang and Leung, 1998).

[c] Expected N_p for the Bayesian optimal design approach (OPT) proposed in this paper.

[d] Maximum sample size per study, based on the table in Wang and Leung's paper.

[e] Probability of a True Positive (PTP) decision for WL and OPT.

[f] Probability of a False Negative (PFN) decision for WL and OPT.

Table 3. Comparison of 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. Treatment-specific success probabilities followed a Beta(0.12, 0.48) distribution.

γ	σ^2 & τ^2	$Mean(\theta_i)$	$\Pr(\theta_i \geq 0.5 \mid \gamma, \tau^2, \sigma^2)$	$\Pr(\theta_i \leq 0.2 \mid \gamma, \tau^2, \sigma^2)$	N_p	No. of Treatments	PTP	PFN
-0.5	0.3^2	0.35	0.284	0.359	31.25	5.93	0.98	0.08
	0.5^2	0.50	0.494	0.176	21.53	6.21	0.98	0.07
	0.8^2	0.64	0.707	0.071	10.37	6.01	0.96	0.08
0	0.3^2	0.32	0.078	0.168	48.15	5.57	0.98	0.05
	0.5^2	0.50	0.501	0.008	27.38	5.39	0.98	0.05
	0.8^2	0.68	0.920	0.000	12.76	5.28	0.95	0.06
0.5	0.3^2	0.31	0.000	0.004	54.05	4.91	0.97	0.03
	0.5^2	0.50	0.500	0.000	31.04	5.01	0.96	0.04
	0.8^2	0.69	1.000	0.000	19.84	4.95	0.94	0.05