Published in final edited form as:

J Biopharm Stat. 2018 ; 28(5): 824–839. doi:10.1080/10543406.2017.1402779.

Bayesian Design for Two-Arm Randomized Phase II Clinical Trials with Endpoints from the Exponential Family Using Multiple Constraints

Wei Jiang*, Jo A. Wick, Jianghua He, Jonathan D. Mahnken, Matthew S. Mayo

Department of Biostatistics, University of Kansas Medical Center, 3901 Rainbow Boulevard 5028 Robinson, Kansas City, Kansas, U.S.A, 66160

Summary:

Frequentist design for two-arm randomized phase II clinical trials with outcomes from the exponential dispersion family was proposed previously, where the total sample sizes are minimized under multiple constraints on the standard errors of the estimated group means and their difference. This design was generalized from an approach specific for dichotomous outcomes. The two previous approaches measure the central tendency of each group and treatment effect based on mean and difference in means. Other measures such as median or hazard ratio are more appropriate under certain situations. In addition, the frequentist approaches assume that unknown parameters are fixed values. This does not reflect the reality that uncertainty always exists for unknowns. Compared to the frequentist methods, the Bayesian approach offers a flexible way to measure central tendency and treatment effect, and incorporate uncertainty in parameters of interest into considerations. In this article, we generalize a Bayesian design for phase II clinical trials with endpoints in the exponential family from the two previously developed frequentist approaches. The proposed design minimizes the total sample sizes under pre-specified constraints on the expected length of posterior credible intervals for measures of treatment effect and central tendency in each group. The design is applicable for trials with fixed or optimal randomization allocation ratio and can be applied under adaptive procedure. Examples of method implementations are provided for different types of endpoints from the exponential family in both fixed and adaptive settings.

Keywords

Multiple constraints; Sample size; Posterior credible interval; Natural conjugate prior family

1 Introduction

Phase II clinical trials play an important role in new treatment development because they provide critical information for the “go or no-go” decision and the design of phase III trials (Yin, 2013). The ability of providing accurate and reliable enough information for the

* willjiang29@gmail.com.

decision making on subsequent phase III trials are especially important and should be the main objective of any phase II programs.

Sample size calculation is important in clinical research because the precision of estimates is directly determined by the size of a study (Piantadosi, 2005). The traditional approaches for determining sample sizes in randomized trials, although vary in specific aspects from each other, mainly aim to provide adequate precision on estimating treatment effect that compares the new therapy to the standard of care or placebo (e.g., hazard ratio and difference in proportions). This is necessary since a phase III program would commence only if a new agent shows sufficient efficacy. However, such demonstrated efficacy depends on information from individual arm and could be misleading when the parameter such as response rate or hazard rate for each arm is poorly estimated. For example, assume that the clinically meaningful difference to be detected is 15 and the variance is 400 for both arms. If we conduct a two-sided normal Z -test at a significance level of $\alpha = 0.05$, the total sample size to achieve a power of 90% is 76, with 38 for each arm. The corresponding margin of error for each arm is therefore $1.96 * \sqrt{400/38} = 6.35$, which provides little protection on the precision for both groups. Now suppose that the true means for the control and novel therapy are 28 and 24.65, respectively. The observed estimates could be potentially 6.35 below and above these two true values, that is 21.65 for the control and 31 for the new treatment. The Z -statistic in this case is equal to 2.04, which falsely recommends proceeding an inactive treatment to phase III. Predicaments like this can be avoided by restricting each arm with a narrower margin of error. Therefore, estimates with adequate precision for both concurrent control and treatment arms should be provided by phase II trials or the conclusions based on demonstrated treatment effects could be misleading. Furthermore, because usually only limited information are available for new therapies, reliable estimates on survival time or response rate, for example, to the new agent are necessary to have a better transition to subsequent phase III trials.

Mayo et al. (2009) developed a two-arm randomized phase II design that allows for estimating the information of interest in each arm and the comparison between two groups with given level of precision. A recent article by Jiang et al. (2016) generalized the work of Mayo et al. (2009) from binary outcome only to endpoints that belong to the exponential dispersion family from a frequentist perspective. Both methods compute the total sample sizes by placing multiple constraints on mean estimates in both control and experiment arms and the estimated difference between means using standard errors. These multiple constraints, alternatively, can be expressed in terms of confidence intervals as $L_\alpha(\mu_C) \leq I_C$, $L_\alpha(\mu_E) \leq I_E$, and $L_\alpha(\mu_E - \mu_C) \leq I$, where $L_\alpha(\cdot)$ represents the length of $100(1 - \alpha)\%$ confidence interval for the parameter of interest, μ_C and μ_E denote the group means, and I_C , I_E , and I are cut-off lengths. Their designs issue more protections on trial estimates and provide more reliable and complete preliminary data to the subsequent phase III programs. Similar to other frequentist methods, however, these two designs assume that the study parameters are fixed values. This does not reflect the realistic situation where uncertainty always exist for unknown parameters and misspecification on these values may lead to poor estimates of total sample sizes. Alternatively, the Bayesian approaches treat the parameters of interest as random variables by nature and aim to take their associated uncertainty into

account. Moreover, while many other measures of central tendency and treatment effect such as median survival and hazard ratio are useful in clinical research, the two previous designs only target on providing precise estimations on group means and the mean difference because the order statistics and ratio problems are mathematically complicated. This is not as troublesome for Bayesian approaches because, with today's computing power, many clever methods such as Markov chain Monte Carlo have been devised for constructing posterior distributions of these parameters. Furthermore, probability in frequentist approaches is defined on the data space and is viewed as the limiting frequency in a large number of trials (Berry et al., 2010). This makes it difficult for clinicians who are less familiar with statistics to understand the interpretation of confidence intervals. In contrast, the Bayesian credible interval is more intuitive and interpretable for clinicians because it views the probability as a degree of belief or knowledge about the unknown (Berry et al., 2010).

Because of the flexibility of Bayesian approaches, Bayesian designs for phase II studies are usually adaptive designs. See Thall and Simon (1994), Stallard (1998), Thall et al. (1995), Stallard et al. (1999), Tan and Machin (2002), Ding et al. (2008), and Zhong et al. (2013) for examples. Although adaptive designs are sometimes more efficient and flexible than traditional single-stage designs, they can be impractical in practice for trials with long treatment duration. They also may increase the risk of operational biases and decrease the accuracy and reliability for assessing the treatment effect (Food and Drug Administration, 2010). The sample size problems assuming a fixed trial size are less frequently discussed for Bayesian approaches in literature. Some examples include Staquet and Sylvester (1977), Sylvester (1988), Brunier and Whitehead (1994), Mayo and Gajewski (2004), and Gajewski and Mayo (2006). Almost all the aforementioned designs are limited to dichotomous outcomes and do not contain a concurrent control arm. Randomized phase II trials are recommended because they provide more accurate and reliable information for the subsequent phase III programs (Rubinstein et al., 2005; Wieand, 2005; Taylor et al., 2006; Rubinstein et al., 2009; Mandrekar and Sargent, 2010). Moreover, many other probability distributions (e.g., exponential, gamma, and normal) from the exponential family are useful in the practice of phase II clinical trials. Some Bayesian sample size computations have been developed for distributions from the exponential family for two-sample comparison. The difference between two binomial proportions was considered by Joseph and Wolfson (1997) based on highest posterior density intervals. Joseph and Belisle (1997) and Pezeshk (2002) applied the Bayesian inference-based approach to determine the sample size for the differences between two normal means. Hand et al. (2011) developed an inferential-theoretic method for sample size determination in the context of comparison of two Poisson rates. However, none of these methods provides protection on the yielding information of new therapy and control arms.

In this article, we extend the work of Mayo et al. (2009) and Jiang et al. (2016) to the entire exponential family from a Bayesian perspective using credible intervals. The proposed Bayesian design is broadly applicable to many commonly used central tendency and treatment effect measurements such as median survival time and hazard ratio. The focus of this article is to provide a fixed pretrial sample size for a two-arm randomized phase II trial. However, the proposed design has the flexibility to be applied in a adaptive framework.

Notice that, for this initial development, we assume that all hyper-parameters are known. An optimal design based on hierarchical models is left for future study. Different from the traditional methods of sample size calculations that only compute a sample size to achieve a desired level of precision for treatment effect, the proposed Bayesian design equips two additional constraints to ensure the reliability and accuracy of information from individual arms. In practice, the precision of estimates of individual arms are seldom considered in designs of randomized phase II trials. We believe that the proposed approach can improve the precision of estimates used for decision making. Our hope is, through our exposition here, to emphasize the importance of these two additional constraints in phase II randomized trials and ultimately improve the success rates of phase III trials.

The remainder of this article is organized as follows. In section 2, we review the average length criterion (ALC). In section 3, the proposed Bayesian design that minimizes the total sample sizes by employing multiple constraints on the average length of posterior credible intervals for the measures of treatment effect and central tendency of both control and treatment arms are derived. The optimal ratio and fixed ratio designs and an adaptive paradigm are explained in this section. We continue in section 4 with three real-life examples of phase II studies to illustrate the application of the proposed Bayesian design in both fixed and adaptive settings. Finally, a brief discussion is given in section 5.

2. Average Length Criterion

The Bayesian criteria of sample size computation generally consist of two classes. One is the decision-theoretic or fully Bayesian approach that maximizes the pre-selected utility function, or equivalently, minimizes the corresponding loss function for determination of sample size. Such approach is less often used in practice because it is practically difficult to develop a widely accepted loss or utility function or a reasonable loss or utility function suitable for the study's specific objectives (Adcock, 1988). The other is the performance-based or inferential-theoretic approach that controls posterior inference for the parameter of interest with certain precision, for which the average coverage criterion (ACC), ALC, and the worst outcome criterion (WOC) are 3 commonly used criteria. We use the ALC criterion introduced by Joseph et al. (1995) because it has practical advantages that neither utility nor loss function needs to be formulated; the ALC criterion is easy to implement and for clinicians to understand; and the ALC criterion computes the sample size by controlling the average length of posterior credible intervals which is, in some sense, comparable to the two original frequentist approaches for which design constraints can be alternatively expressed in terms of confidence intervals. However, we do not advocate for using one approach over the other.

Suppose that $\mathbf{y} = (y_1, \dots, y_n)$ is a vector of n independent and identically distributed data points drawn from a population indexed by $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$. The parameter of interest is a function of $\boldsymbol{\theta}$ denoted by $\mu = \mu(\boldsymbol{\theta})$ with dimension one. If the sampling distribution is $p(\mathbf{y}|\boldsymbol{\theta})$, and $p(\boldsymbol{\theta})$ is the chosen prior density, then the posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}.$$

In practice, the marginal posterior density for $\mu = \mu(\boldsymbol{\theta})$, denoted by $p(\mu|\mathbf{y})$, can be computed from $p(\boldsymbol{\theta}|\mathbf{y})$ using simulations.

The ALC criterion fixes α and determines the smallest n by guaranteeing that the expected length of a $100(1 - \alpha)\%$ posterior credible interval for the parameter of interest μ weighted by the marginal $p(\mathbf{y})$ does not exceed a pre-specified length l . Mathematically, the minimum n is chosen such that

$$ALC_{\alpha}(\mu|n) = \int \dots \int l_{\alpha}(\mu|\mathbf{y}, n) p(\mathbf{y}) d\mathbf{y} \leq l, \quad (1)$$

where $l_{\alpha}(\mu|\mathbf{y}, n)$ is the length of the $100(1 - \alpha)\%$ posterior credible interval of μ given data $\mathbf{y}$ with size n . If we draw a large sample $\{(\mathbf{y}_s, \mu_s), s = 1, \dots, S\}$ from the joint distribution $p(\mu, \mathbf{y})$, the integral (1) can be estimated using Monte Carlo integration as

$$\widehat{ALC}_{\alpha}(\mu|n) = \frac{1}{S} \sum_{s=1}^S l_{\alpha}^s(\mu|\mathbf{y}_s, n). \quad (2)$$

In this article, the lengths of credible intervals are calculated based on central posterior interval due to its easy computation and direct interpretation as the posterior $\alpha/2$ to $1 - \alpha/2$ inter-quantile range.

3. Bayesian Design with Multiple Constraints

We start this section by refreshing the association between the exponential family and the natural conjugate prior distributions.

3.1 Exponential family and natural conjugate prior

A conjugate model is usually a good starting point for Bayesian analysis since it simplifies the computation and is easy to understand. The review of the natural conjugate distributions and the exponential family closely follow Gelman et al. (2014).

Suppose that $\mathbf{y} = (y_1, \dots, y_n)$ are n independent and identically distributed samples from an exponential family, the likelihood function is then

$$p(\mathbf{y}|\boldsymbol{\theta}) = \left\{ \prod_{i=1}^n h(y_i) \right\} \{c(\boldsymbol{\theta})\}^n \exp \left\{ \sum_{j=1}^k \left[\eta_j(\boldsymbol{\theta}) \sum_{i=1}^n t_j(y_i) \right] \right\}, \quad (3)$$

where the statistic $T(\mathbf{y}) = (\sum_{i=1}^n t_1(y_i), \dots, \sum_{i=1}^n t_k(y_i))$ is a sufficient statistic for $\boldsymbol{\theta}$, and the vector $\boldsymbol{\eta} = (\eta_1(\boldsymbol{\theta}), \dots, \eta_k(\boldsymbol{\theta}))$ is called the natural parameter. The densities of natural conjugate prior family have the same functional form as the likelihood function (Gelman et al., 2014):

$$p(\boldsymbol{\theta}) \propto \{c(\boldsymbol{\theta})\}^{\tau_0} \exp \left\{ \sum_{j=1}^k \eta_j(\boldsymbol{\theta}) \tau_j \right\} \quad (4)$$

where $(\tau_0, \dots, \tau_k)$ are hyper-parameters. The posterior density for θ is then

$$p(\theta | \mathbf{y}) \propto \{c(\theta)\}^{\tau_0 + n} \exp \left\{ \sum_{j=1}^k \eta_j(\theta) \left[\tau_j + \sum_{i=1}^n t_j(y_i) \right] \right\}. \quad (5)$$

Notice that the hyper-parameters can be viewed as prior observations. In the following three sections, we explain the proposed Bayesian design for two-arm, randomized phase II clinical trials. It minimizes the total sample sizes using multiple constraints subject to ALC criterion. The design allows to obtain estimates of central tendency in two individual arms with given levels of precision and also protects the estimate of treatment effect within certain predefined bounds. This setting should improve the completeness, accuracy and reliability of information provided from phase II programs.

3.2 Design constraints

Let $\mathbf{y}_C = (y_{1,C}, \dots, y_{n_C,C})$ and $\mathbf{y}_E = (y_{1,E}, \dots, y_{n_E,E})$ denote n_C and n_E independent and identically distributed draws from two independent exponential family distributions for the control and experiment arms, respectively. We assume that joint prior distribution $p(\theta_C, \theta_E)$ is equal to $p(\theta_C)p(\theta_E)$; that is, θ_C and θ_E are independent of each other. We are interested in minimizing the total sample size subject to the expected length of posterior credible intervals for the measures of central tendency and treatment effect; namely, the parameters of primary interest are $\mu_C = \mu_C(\theta_C)$, $\mu_E = \mu_E(\theta_E)$, and $\delta = \delta[\mu_C(\theta_C), \mu_E(\theta_E)]$. Using equation (1) in section 2, the total sample size, defined as $N = n_C + n_E$, is minimized under the following three constraints:

$$ALC_\alpha(\mu_C | n_C) = \int \dots \int l_\alpha(\mu_C | \mathbf{y}_C, n_C) p(\mathbf{y}_C) d\mathbf{y}_C \leq l_C, \quad (6)$$

$$ALC_\alpha(\mu_E | n_E) = \int \dots \int l_\alpha(\mu_E | \mathbf{y}_E, n_E) p(\mathbf{y}_E) d\mathbf{y}_E \leq l_E, \quad (7)$$

$$ALC_\alpha(\delta | n_C, n_E) = \int \dots \int l_\alpha(\delta | \mathbf{y}_C, \mathbf{y}_E, n_C, n_E) p(\mathbf{y}_C, \mathbf{y}_E) d\mathbf{y}_C d\mathbf{y}_E \leq l_\delta. \quad (8)$$

Performing direct calculations on the above three integrals seems intractable; therefore, with a pre-specified α level and the hyper-parameters $(\tau_{0,C}, \dots, \tau_{k_C,C})$ and $(\tau_{0,E}, \dots, \tau_{k_E,E})$ for two arms, the following Monte Carlo algorithm is applied to identify n_C and n_E that satisfy constraints (6) – (8).

1. For $s = 1, \dots, S$,
 - a. Draw θ_C^s and θ_E^s from $p(\theta_C)$ and $p(\theta_E)$, respectively; then draw $\mathbf{y}_C^s$ and $\mathbf{y}_E^s$ from $p(\mathbf{y}_C | \theta_C^s)$ and $p(\mathbf{y}_E | \theta_E^s)$, respectively.

- b. Generate M draws of θ_C and θ_E from $p(\theta_C|y_C^s)$ and $p(\theta_E|y_E^s)$, respectively; then compute $\mu_C(\theta_C)$, $\mu_E(\theta_E)$, and $\delta[\mu_C(\theta_C), \mu_E(\theta_E)]$ for each draw.
 - c. Obtain length of credible intervals $l_\alpha^s(\mu_C|y_C^s, n_C)$, and $l_\alpha^s(\mu_E|y_E^s, n_E)$, and $l_\alpha^s(\delta|y_C^s, y_E^s, n_C, n_E)$ by taking the $\alpha/2$ and $1 - \alpha/2$ sample quantiles from M values of μ_C , μ_E , and $\delta(\mu_C, \mu_E)$ derived in step 2.
 2. Left-hand sides of constraints (6) – (8) can be approximated using formula (2) in section 2 as

$$\begin{aligned}\widehat{ALC}_\alpha(\mu_C|n_C) &= \frac{1}{S} \sum_{s=1}^S l_\alpha^s(\mu_C|y_C^s, n_C). \\ \widehat{ALC}_\alpha(\mu_E|n_E) &= \frac{1}{S} \sum_{s=1}^S l_\alpha^s(\mu_E|y_E^s, n_E). \\ \widehat{ALC}_\alpha(\delta|n_C, n_E) &= \frac{1}{S} \sum_{s=1}^S l_\alpha^s(\delta|y_C^s, y_E^s, n_C, n_E).\end{aligned}$$

The above Monte Carlo procedure computes the lengths of posterior credible intervals based on M posterior samples of two measures of central tendency and the corresponding parameter of treatment effect with each credible interval been evaluated S times. Fairly good accuracy should be retained as long as the selected posterior and Monte Carlo sample sizes are large enough. A simulation study is carried out to guide the selection of M and S assuming the prior and likelihood specifications are appropriate. It can be seen in Figure 1 that, the center of estimated one-to-one ratio total sample size for an exponential outcome using median as the measure of central tendency and hazard ratio as the measure of treatment effect is fairly stable as the number of independent runs (i.e., $S = M$) increases. The trial-to-trial variability is relatively large for small $S = M$. It is fairly stable and reaches a reasonably small magnitude when $S = M \geq 3000$. This suggests that the Monte Carlo approach may work well for the ALC criterion with relatively small S and M . One could further reduce Monte Carlo errors by averaging over a fixed number of parallel simulation trains.

3.3 Fixed ratio design

Sometimes investigators desire to assign patients to the control and experiment arms with a fixed ratio $r = c/e$, where c and e are both positive integers. For example, economic efficiency can potentially be improved by assigning more patients to the less expensive arm. When this is the case, the desired point (n_C, n_E) needs to satisfy the relationship $n_C = mn_E$ as well as constraints (6) – (8). Under the assumption that the greatest common divisor for c and e is one (i.e., relative prime), Mayo et al. (2009) demonstrated that the set $\{(cd, ed) : d = 1, 2, 3, \dots, \infty\}$ contains all two-dimensional points such that their coordinates satisfy the ratio $r = c/e$.

We perform the method of grid search to identify the minimum sample size that satisfies design constraints and the fixed ratio $r = c/e$. To begin with, the crude starting value of d can be estimated by applying the Monte Carlo simulation algorithm discussed in section 3.2 with a small value of $S = M$ (say, 500). Starting from grid value $d = 1$, we first substitute (cd, ed) for (n_C, n_E) in three integrals under constraints (6) – (8), and then apply the Monte Carlo algorithm to evaluate each constraint. If any of them fail to meet the corresponding constraint (or constraints), we increase d by 1. This procedure is not terminated until all estimated expected lengths of posterior credible intervals are smaller than or equal to pre-specified upper limits. The resulting new d (i.e., d_0) is the estimated starting value. Now we increase the number of independent runs to a larger value ($S = M \geq 3000$) to reduce trial-to-trial sample size variability. The fixed ratio minimum sample size is derived by applying the following algorithm:

Step 1: Let $d = d_0$ be the starting grid value.

Step 2: With defined c and e , replace (n_C, n_E) in the left-hand side of constraints (6) – (8) by (cd, ed) . Then, apply the Monte Carlo algorithm to approximate integrals under three design constraints. If any lead to values larger than the corresponding pre-specified cutoff levels (i.e., I_C , I_E and I_δ), then $d = d + 1$; Otherwise, using $d = d - 1$ to evaluate whether current d value is actually larger than the desired minimum.

Step 3: Repeat step 3 until the minimum d is attained. The minimum sample size for $c : e$ fixed ratio design is $(n_C, n_E) = (cd, ed)$.

3.4 Optimal ratio design

Without limitation on allocation ratio, the minimum sample size only needs to satisfy constraints (6) – (8). The optimal ratio design has two-dimensional grid points (i.e., (n_C, n_E)). The standard grid search method that evaluates all combinations of grid values of n_C and n_E is inefficient and extremely time consuming. As the ranges of n_C and n_E increase, the number of grid points inflates considerably. For example, when both n_C and n_E range from 1 to 10 by increments of 1, the total number of grids is $10^2 = 100$. If both n_C and n_E take integer values from 1 to 30, the total number of grids overwhelmingly increases to $30^2 = 900$. In this section, we propose a rapid grid search method that lightens the computational burden by efficiently reducing the number of grid points.

Figure 2 displays the rapid grid search (RGS) procedure. To begin with, the standard grid search method is conducted to obtain the minimal sample size $(n_C^{(6)}, n_E^{(7)})$ under constraints (6) and (7). We set $(n_C^0, n_E^0) = (n_C^{(6)}, n_E^{(7)})$ as the starting grid point of RGS. Be aware that $(n_C^{(6)}, n_E^{(7)})$ reaches the minimal sample size if it does not violate constraint (8). Then we add one subject at a time to both control and experiment arms and evaluate constraint (8) for each grid point until the approximated expected length of the posterior credible interval for $\delta(\mu_C, \mu_E)$ is smaller than I_δ . This gives us an initial optimal candidate (n_C^1, n_E^1) . The subsequent searches proceed in two directions. At one direction, the n_E^1 remains unchanged but one subject at a time is removed from the control arm until the grid point fails to meet

constraint (8). The last grid point that satisfies constraint (8) is the second optimal candidate (n_C^2, n_E^2) . It next adds one subject to n_E^2 and decreases one subject from n_C^2 and then makes subsequent searches towards left by withdrawing one subject at a time from the control arm until the grid point fails to satisfy constraint (8). Suppose that k optimal Candidates $\{(n_C^j, n_E^j), j = 1, \dots, k\}$ are obtained by repeating the above steps, this procedure is terminated at this direction if $n_C^k - 1$ is smaller than n_C^0 . The RGS procedure at the other direction follows the same strategy. We give the algorithm of the proposed RGS method as follows:

Step 1: use the Monte Carlo approximation and the standard grid search method to derive the minimum sample size under constraints (6) and (7). Consider this point as the starting grid point and denote it by (n_C^0, n_E^0) . If $\widehat{ALC}_\alpha(\delta | n_C^0, n_E^0) \leq l_\delta$, then (n_C^0, n_E^0) is the optimal point under the optimal ratio design; otherwise, go to step 2;

Step 2: For $i = 1, 2, \dots, \infty$, evaluate $\widehat{ALC}_\alpha(\delta | n_C^0 + i, n_E^0 + i)$ until it is no greater than l_δ . suppose $i = I$, then report $(n_C^{opt}, n_E^{opt}) = (n_C^0 + I, n_E^0 + I)$ as an optimal candidate and proceed to step 3;

Step 3:

Direction 1: for $m = 0, 1, \dots, n_C^{opt}$, evaluate $\widehat{ALC}_\alpha(\delta | n_C^{opt} - m, n_E^{opt})$. Find the largest value $m = m'$ that satisfies constraint (8). Update $(n_C^{opt}, n_E^{opt}) = (n_C^{opt} - m', n_E^{opt})$ and report this point as an optimal candidate. Proceed to direction 1 at step 4;

Direction 2: for $m = 0, 1, \dots, n_E^{opt}$, evaluate $\widehat{ALC}_\alpha(\delta | n_C^{opt}, n_E^{opt} - m)$ Find the largest value $m = m^*$ that satisfies constraint (8). Update $(n_C^{opt}, n_E^{opt}) = (n_C^{opt}, n_E^{opt} - m^*)$ and report this point as an optimal candidate. Proceed to direction 2 at step 4;

Step 4:

Direction 1: Stop the procedure at direction 1 if $n_C^{opt} - 1 < n_C^0$; otherwise update $(n_C^{opt}, n_E^{opt}) = (n_C^{opt} - 1, n_E^{opt} + 1)$ and repeat direction 1 at step 3;

Direction 2: Stop the procedure at direction 2 if $n_E^{opt} - 1 < n_E^0$; otherwise update $(n_C^{opt}, n_E^{opt}) = (n_C^{opt} + 1, n_E^{opt} - 1)$ and repeat direction 2 at step 3;

Step 5: Output those optimal candidates with the smallest total sample size.

For the optimal ratio design, the procedure of RGS may identify multiple optimal points with the same minimum total sample size but different allocation schemes. During clinical practice, investigators can choose the point that most satisfies their needs. For example, a design that assigns more patients to the less expensive arm may be preferred for the purpose of budgeting.

3.5 Adaptive paradigm

The bulk of this article is devoted to describing the minimum pretrial sample size determination using the proposed Bayesian design for two-arm, randomized phase II trials. We pause briefly here to provide a paradigm explaining how the proposed Bayesian design can be applied in the adaptive settings.

An appealing advantage of Bayesian approaches is the flexibility of continually updating Bayesian inferences and sample sizes as data accumulate. Because the precision of estimation is a direct consequence of the size of a study (Piantadosi, 2005), re-estimating sample size based on most current information at an interim stage may improve the accuracy and reliability of the resulting estimates. One may consider discontinuing a trial early if, at an interim stage, $Pr(\delta > \epsilon | data) > 1 - c$ or $Pr(\delta < \epsilon | data) > 1 - c$. This is equivalent to stopping when a $1 - 2c$ posterior credible interval excludes. Different c values may be used for efficacy and futility. Rules similar to this have been proposed by many authors including Mehta and Cain (1984), Berry (1985), Freedman and Spiegelhalter (1989), and Thall and Simon (1994). If the trial is not discontinued early, the sample size needed for providing precise estimates for measures of central tendency and treatment effect can be re-estimated using the posterior distributions based upon current accumulated data as the updated new priors. Therefore, one way to apply the proposed Bayesian design in the adaptive framework would be:

1. Suppose the data collected at the interim stage from both control and treatment arms are y_C^I and y_E^I , respectively. The trial is stopped early due to efficacy or futility if $Pr[\delta(\mu_C, \mu_E) > \epsilon | y_C^I, y_E^I] > 1 - c_I$ or $Pr[\delta(\mu_C, \mu_E) < \epsilon | y_C^I, y_E^I] > 1 - c_I$, where c_I represents the threshold for interim decision rules. The trial continues otherwise.
2. If the trial is not stopped early, the posterior distributions based on current accumulated data are employed as new priors to re-estimate the sample size.
3. If the data observed at final stage are $y^F = (y_C^F, y_E^F)$, evaluate the treatment effect based on $Pr[\delta(\mu_C, \mu_E) > \epsilon | y_C^F, y_E^F] > 1 - c_F$ and $Pr[\delta(\mu_C, \mu_E) < \epsilon | y_C^F, y_E^F] > 1 - c_F$, where c_F represents the threshold for final decision rules.

Although it is not required for the Bayesian procedure to evaluate the frequentist operating characteristics, the frequentist soundness such as type I error rate of this procedure can be checked and controlled by calibrating the threshold parameters c_I and c_F .

3.6 Tuning parameters and prior elicitation

Similar to other methods of sample size determination, the tuning cutoffs that need to be pre-specified for constraints (6) – (8) affect the final minimum sample sizes. Inappropriate large cutoffs may lead to inadequately sized trials and poor estimations, whereas unnecessarily small cutoffs may result in lengthy and expensive trials. In practice, selecting appropriate tuning cutoffs for design constraints is not an easy task. It requires close cooperation and communication between statisticians and clinicians. For example, if

clinicians believe that the mean duration of a symptom should be varied no greater than a radius of 10 weeks, then 20 weeks can be selected for the upper cutoff.

Prior distributions need to properly reflect experts' opinion. This again requires close cooperation and communication between statisticians and clinicians. For example, the center value of the chosen prior can be elicited by asking clinicians questions such as, "What is the most likely average value that you expect to observe for the primary endpoint?" The response to this question could correspond to the prior mean, median and mode. Likewise, the uncertainty in the center value can be elicited by asking clinicians questions such as, "How uncertain is the center value you provided?" The response to this question could be associated with prior sample size or the width of a $100(1 - \alpha)\%$ credible interval of the prior distribution (Mayo and Gajewski, 2004). The computer plot of the pdf or pmf of the corresponding conjugate distribution may be helpful for clinicians to conceptualize the prior (Thall and Simon, 1994). For exponential family, hyper-parameters of natural conjugate priors can be interpreted as additional observations (Gelman et al., 2014). If an estimated total sample size for an experiment is N and n_0 prior observations are assumed, the ratio $n_0/(n_0 + N)$, in some sense, can be viewed as the percentage of information contributed by the prior distribution.

4. Illustrative Examples

In section 3, the proposed Bayesian design for two-arm randomized phase II clinical trials was discussed. In this section, we present in sections 4.1 and 4.2 with two real-life phase II clinical trials to demonstrate the application of the proposed Bayesian design on normal and Bernoulli distributed endpoints. Moreover, a breast cancer study is used to demonstrate the application of the proposed approach on exponential endpoint in both fixed and adaptive settings. R functions are developed and used for computing sample sizes in all examples.

4.1 Normal endpoint

The normal distribution is often assumed when the endpoint is continuous. It is useful in phase II trials when the original values used for categorization are of interest. For example, instead of dichotomizing the change of tumor size in terms of RECIST criteria (Eisenhauer et al., 2009), the change of tumor size itself is sometimes used as the primary endpoint. Lavin (1981) demonstrated that the log of tumor size ratio to baseline approximately followed the normal distribution.

4.1.1 Normal-Inverse χ^2 model.—Under the two-parameter normal sampling model, we have $\theta_g = (\mu_g, \sigma_g^2)$, $g = C, E$. Recall from equation (4) that the natural conjugate prior for θ_g has the same functional form as the likelihood. After some algebra, it can be shown that

$$p(\theta_g) \propto \left(\frac{1}{\sigma_g^2}\right)^{\frac{v_g}{2} + 1} \exp\left(-\frac{v_g \omega_g^2}{2\sigma_g^2}\right) \left(\frac{1}{\sigma_g^2}\right)^{\frac{1}{2}} \exp\left(-\frac{\kappa_g}{2\sigma_g^2}(\mu_g - \zeta_g)^2\right).$$

The corresponding conjugate prior density for $\theta_g = (\mu_g, \sigma_g^2)$ is a mixture of the marginal prior distribution $\sigma_g^2 \sim \text{Inverse} - \chi^2(v_g, \omega_g^2)$ and the conditional prior distribution $\mu_g | \sigma_g^2 \sim N(\zeta_g, \sigma_g^2/\kappa_g)$. We label this the Normal-Inv $\chi^2(\zeta_g, \sigma_g^2/\kappa_g; v_g, \omega_g^2)$, where ζ_g and σ_g^2/κ_g represent location and scale of μ_g with κ_g prior measurements, and $v_g \omega_g^2$ can be viewed as prior sum of squares (Gelman et al., 2014). After some algebraic manipulation, it can be shown that the joint posterior density for $\theta_g = (\mu_g, \sigma_g^2)$ is

$$\theta_g | y_g \sim \text{Normal} - \text{Inv} \chi^2 \left(\zeta_{n,g}, \frac{\omega_{n,g}^2}{\kappa_{n,g}}; v_{n,g}, \omega_{n,g}^2 \right),$$

where

$$\begin{aligned} \zeta_{n,g} &= \frac{\kappa_g}{\kappa_a + n_a} \zeta_g + \frac{n_g}{\kappa_a + n_a} \bar{y}_g, \quad \kappa_{n,g} = \kappa_g + n_g, \quad v_{n,g} = v_g + n_g, \quad \text{and} \\ v_{n,g} \omega_{n,g}^2 &= v_g \omega_g^2 + \sum_{i=1}^{n_g} (y_{i,g} - \bar{y}_g)^2 + \frac{\kappa_g n_g}{\kappa_g + n_g} (\bar{y}_g - \zeta_g)^2. \end{aligned}$$

Notice that both the prior information and the information provided by the data are contained in the parameters of the joint posterior distribution. For instance, the parameter $\zeta_{n,g}$ can be explained as a weighted average of the prior mean ζ_g and the sample mean $\bar{y}_g$ (Gelman et al., 2014).

4.1.2 Illustrative example: change in tumor size as endpoint.—A published randomized phase II clinical study of the effect of sorafenib and erlotinib in non-small-cell lung cancer (Karrison et al., 2007) treats tumor size as a continuous variable and aims to assess antitumor activity. The control arm of this study was 150 mg erlotinib daily plus placebo, and the two treatment arms were 150 mg erlotinib daily plus 200 mg and 400 mg sorafenib twice daily, respectively. The ratio of tumor size at 8 months after treatment to that at baseline was the primary endpoint, and the author assumed that the log of this ratio came from a normal distribution. In this example, we use the control arm and the 150 mg erlotinib plus 400 mg sorafenib to illustrate the proposed design. Based upon four previous single-arm studies of erlotinib or sorafenib, this study assumed a mean log ratio of 0.05 for the placebo arm and of -0.13 for the chosen experiment arm. In addition, this trial assumed an equal variance of 0.1197 for the log ratio by pooling the variance estimates from the four previous trials. For more details about this phase II cancer trial, see Karrison et al. (2007).

In this study, the mean and difference in means were used to measure the central tendency and treatment effect. Because the mean parameter of a normally distributed endpoint is $\mu_g(\theta_g) = \mu_g$, $g = C, E$, the design constraints are $ALC_a(\mu_C | n_C) \leq I_C$, $ALC_a(\mu_E | n_E) \leq I_E$, and $ALC_a(\delta | n_C, n_E) = ALC_a(\mu_E - \mu_C | n_C, n_E) \leq I_\delta$. Suppose that the appropriate choices for cutoff levels of design constraints are $I_C = I_E = 0.2$ and $I_\delta = 0.28$, this would ensure that, given the computed sample sizes, the 95% posterior credible intervals for μ_C and μ_E on average, are no wider than 0.2 and the average length of credible interval for $\mu_E - \mu_C$ is no

greater than 0.28. For prior elicitation, we assume that $E(\mu_C|\sigma_C^2) = \zeta_C = 0.05$ for the control arm and $E(\mu_E|\sigma_E^2) = \zeta_E = -0.13$ for the experiment arm. We elicit the prior center values for σ_C^2 and σ_E^2 based on mode due to the high skewness of the prior conjugate distributions for both σ s. Namely, we assume $\text{Mode}(\sigma_C^2) = (v_C \omega_C^2)/(v_C + 2) = 0.1197$ and $\text{Mode}(\sigma_E^2) = (v_E \omega_E^2)/(v_E + 2) = 0.1197$. Moreover, we let $v_C = v_E = 10$ so relatively small information for σ s are provided by the priors. For comparative purposes, we report the total sample sizes using both weekly informative and informative priors for mean parameters μ_C and μ_E . We let $\kappa_C = \kappa_E = 0.01$ for weakly informative prior, namely, nearly no prior information is provided for μ_C and μ_E . In this case, the optimal ratio design requires enrolling 136 total subjects where allocation schemes include (65, 71), (66, 70), (67, 69), (68, 68), (69, 67), (70, 66), and (71, 65); the one-to-one design leads to 136 total subjects among which each arm contains 68 subjects; and the one-to-two design results in 195 total subjects where 65 individuals are assigned to the control and 130 individuals are assigned to the experiment arm. On the contrary, an informative model with $\kappa_C = \kappa_E = 30$ results in 77, 78, and 108 total subjects for the optimal, one-to-one, and one-to-two ratios, respectively. The optimal ratio design using informative priors also yields multiple allocation schemes including (36, 41), (37, 40), (38, 39), (39, 38), (40, 37), and (41, 36). As expected, the total sample sizes produced under weakly informative model are larger than the ones obtained from informative model. Further, it should be noticed that the resulting total sample size for Normal-Inverse χ^2 model is irrelevant to the location of center value of μ_C and μ_E but affected by the uncertainty in the center values. This is because the variance and mean of the normal distribution are independent of each other.

In addition, we compare the proposed Bayesian design to the frequentist method from Jiang et al. (2016). The design by Jiang et al. (2016) intends to identify the minimal sample sizes which ensure the 95% confidence intervals for center parameters in both control and experiment arms and the parameter of treatment effect to be no wider than I_C , I_E , and I_δ , respectively. Based on the same assumptions and tuning parameters, the frequentist design requires 94, 94 and 138 subjects for the optimal, one-to-one, and one-to-two ratio designs. Like the Bayesian approach, the frequentist method produces several allocation scenarios for the optimal ratio design including (48, 46), (47, 47), and (46, 48). It can be observed that the sample sizes yielded from the frequentist approach are smaller than the ones obtained from the weakly informative model but are larger than the ones based on the informative priors for all one-to-one, one-to-two, and optimal ratio designs.

4.2 Bernoulli endpoint

Binary outcomes, such as whether a partial or complete tumor response is observed from a patient or whether a patient has survived one year later after receiving treatment, are commonly encountered in phase II oncology trials.

4.2.1 Beta-Bernoulli model.—The natural conjugate prior for Bernoulli distribution follows beta distribution with parameters α_g and β_g , $g = C, E$. That is

$$p(\theta_g) \propto \theta_g^{\alpha_g - 1} (1 - \theta_g)^{\beta_g - 1}.$$

Using equation (5), the corresponding posterior distribution for θ_g is therefore

$$\theta_g \sim \text{Beta}\left(\alpha_g + \sum_{i=1}^{n_g} y_{i,g}, \beta_g + n_g - \sum_{i=1}^{n_g} y_{i,g}\right).$$

Be aware that the above two hyper-parameters can be explained as $\alpha_g - 1$ prior successes and $\beta_g - 1$ prior failures for a particular sequence of $\alpha_g + \beta_g - 2$ prior experiments (Gelman et al., 2014).

4.2.2 Illustrative example: tumor response as endpoint.—A phase II clinical trial at the University of Kansas Medical Center assessed and compared the response rates of two chemotherapy regimens in non-small-cell lung cancer. The standard of care of this study was pemetrexed and cisplatin, and the experiment arm was the combination of pemetrexed/cisplatin with MK-0646. This trial used the frequentist design developed by Mayo et al. (2009) to compute the total sample size. It was designed to make the length of a 90% confidence interval for the response rate in each arm no wider than 0.33, and for the difference between the response rates no greater than 0.4. The expected response rate of 0.3 was used for the control arm, and 0.45 was the expected rate for the experiment arm. As a result, the optimal, one-to-one, and one-to-two ratios lead to 62, 62, and 69 total patients, respectively.

The parameters of interest in this study are the response rates of each group (i.e., $\mu_g(\theta_g) = \theta_g$, $g = C, E$) and the difference in response rates (i.e., $\theta_E - \theta_C$). Therefore, the three constraints under which the total sample size is minimized are $ALC_a(\theta_C | n_C) \leq I_C$, $ALC_a(\theta_E | n_E) \leq I_E$, and $ALC_a(\theta_E - \theta_C | n_C, n_E) = ALC_a(\theta_E - \theta_C | n_C, n_E) \leq I_\delta$. Using the prior information provided above, we let $E(\theta_C) = \alpha_C / (\alpha_C + \beta_C) = 0.3$ and $E(\theta_E) = \alpha_E / (\alpha_E + \beta_E) = 0.45$. Suppose that we have a strong belief in the prior knowledge about θ_C (i.e., standard care) but are not confident in the prior information for θ_E (i.e., new treatment). This may be reflected by placing an informative prior on the control arm but a weakly informative prior on the experiment arm. More specifically, we let $\alpha_C + \beta_C = 12$ which corresponds to 10 prior observations, and $\alpha_E + \beta_E = 2$ which corresponds to no prior observation. Using the same cutoff levels of $I_C = I_E = 0.33$ and $I_\delta = 0.4$ as those selected in the frequentist method, the optimal ratio design leads to 34 total patients; the one-to-one design requires 36 total patients, and the one-to-two design results in 36 total patients where 12 individuals are assigned to the control and 24 individuals are assigned to the experiment arm. Notice that multiple allocation schemes are produced by the optimal ratio design including (12, 22), (13, 21), (14, 20), and (15, 19). As expected, the sample sizes based on the frequentist approach are larger than the ones based on Bayesian approach because strong prior information is assumed for the control arm.

4.3 Exponential endpoint

The exponential distribution is commonly used in phase II clinical trials when the primary endpoint is a time-to-event outcome such as progression-free survival or overall survival.

4.3.1 InverseGamma-Exponential model.—The natural conjugate prior for exponential distribution is

$$p(\theta_g) \propto \left(\frac{1}{\theta_g}\right)^{\alpha_g+1} \exp\left\{-\frac{\beta_g}{\theta_g}\right\}, \quad g = C, E,$$

where $\tau_{0,g} = \alpha_g + 1$ and $\tau_{1,g} = \beta_g$. This is an Inverse-Gamma density with shape parameter α_g and rate parameter β_g . The corresponding posterior distribution for θ_g is

$$\theta_g | y_g \sim \text{Inverse-Gamma}\left(\alpha_g + n_g, \beta_g + \sum_{i=1}^{n_g} y_{i,g}\right).$$

The above two hyper-parameters can be viewed as $\alpha_g + 1$ prior observations with total waiting time β_g (Gelman et al., 2014).

4.3.2 Illustrative example: overall survival as endpoint.—The primary objective of a two-arm, randomized, phase II breast cancer study was to evaluate and compare the overall survival of a new therapy versus a standard of care. The trial was planned based on the assumptions of a median overall survival of 15 months for the control arm and a hazard ratio of 0.6, so the assumed median overall survival for the experiment arm was 25 months.

Recall that the median of an $\exp(\theta_g)$ outcome is $\mu_g(\theta_g) = \log(2)\theta_g$; therefore, the three design constraints under which the total sample size is minimized are: $ALC_a[\log(2)\theta_C | n_C] \leq I_C$, $ALC_a[\log(2)\theta_E | n_E] \leq I_E$, and $ALC_a[\delta | n_C, n_E] = ALC_a[\log(\theta_C/\theta_E) | n_C, n_E] \leq I_\delta$. We consider mode to elicit the four hyper-parameters α_C , β_C , α_E , and β_E due to the skewness of Inverse-Gamma priors. Based on the design assumptions of this trial, we assume $Mode(\theta_C) = \beta_C/(\alpha_C + 1) = 15/\log(2) = 21.6$ for the control and $Mode(\theta_E) = \beta_E/(\alpha_E + 1) = 25/\log(2) = 36.1$ for the new regimen. For illustration purposes, we also compute the sample sizes when the assumed hazard ratio is 0.5, for which the corresponding prior center value for θ_E is 43.3. In order to study the influence of prior specification on sample size estimates, we place both weakly informative and informative priors on θ_C and θ_E . More specifically, the prior sample sizes of both arms are 10 for the weakly informative model, and are 20 for the informative model. The cutoff values used in this example are $I_C = I_E = 13.5$ and $I_\delta = 0.65$. The computed minimum sample sizes under the one-to-one, one-to-two, and optimal ratios are reported in Table 1. Noted that, except for the one-to-two ratio with informative priors, computed sample sizes for all other scenarios assuming a hazard ratio of 0.5 are larger than the ones assuming a hazard ratio of 0.6. This is different from the traditional frequentist parametric approach of sample size calculation for exponential distribution where larger effect size usually requires a smaller sample size. This is because the traditional method only focuses on detecting a treatment difference but issues no control on precision of estimates

for two individual arms. Recall that the variance of an exponentially distributed variable is equal to the square of its mean. This means that as the difference between treatment and control becomes larger, the variance in experiment arm also increases. In this case, more subjects may be needed to protect the estimate of new therapy to be within a certain predefined bound. This explains the importance of the proposed two additional constraints on both individual arms because they are the “firewall” to prevent poor and unreliable estimations, the major causal factor that leads to failure in phase III studies.

The application of the proposed Bayesian approach in adaptive settings is also explored in this section using the breast cancer example. Now suppose that reliable information are available for the standard of care but not for the new therapy. This can be reflected by employing a larger prior sample size of 20 for the control arm but a smaller prior sample size of 10 for the experiment arm. The prior assumptions are $Mode(\theta_C) = 21.6$, $Mode(\theta_E) = 36.1$, $a_C = 19$, and $a_E = 9$. We assume an equal randomization and consider an interim analysis when 50% observations are available in this example. Using the same tuning parameters of $I_C = I_E = 13.5$, and $I_\delta = 0.65$, this gives us a total sample size of 154 (77 in each arm). In order to control the frequentist soundness, we conduct a simulation study based on $S = M = 3000$ and 10000 iterations showing that $1 - c_I = 0.99$ and $1 - c_F = 0.98$ ensure a two-sided type I error rate of 5%. Suppose that at the interim stage (at which 38 observations for each arm are available) the observed median survival for control and experiment arms are 14.3 months and 14.7 months, the interim decision rules of

$Pr[\log(\theta_C/\theta_E) > 0 | y_C^I, y_E^I] = 45.1\% < 1 - c_I$ and $Pr[\log(\theta_C/\theta_E) < 0 | y_C^I, y_E^I] = 54.9\% < 1 - c_I$ suggest continuing the study. Based on the current available information, the updated total sample size for the remaining of the trial is 44 (22 in each arm). This allows us to save some subjects thus money for a less promising trial but, importantly, still preserves the reliability of estimations. Furthermore, if the median survival times observed at the interim stage are 26.4 and 39.1, both of which are much longer than the original assumptions. In this case, the interim decision rules are $Pr[\log(\theta_C/\theta_E) > 0 | y_C^I, y_E^I] = 1.4\% < 1 - c_I$ and

$Pr[\log(\theta_C/\theta_E) < 0 | y_C^I, y_E^I] = 98.6\% < 1 - c_I$, suggesting to proceed to final stage. Based on current data, a total of 108 subjects is needed for the second part of the trial, which is larger than the original value of 78. This is a result of larger median survival times thus larger variances shown from the current data for both control and treatment.

5. Discussion

In this article, we developed a Bayesian design and discussed its applications in two-arm randomized phase II clinical trials with endpoints from exponential family. The proposed design minimizes the sample sizes by placing multiple constraints on the average length of posterior credible intervals for measures of treatment effect and central tendency of individual arms. This is important for designing phase II clinical trials since they ensure the sufficiency and reliability of the resulting estimations. These information are critical for the future phase III studies because they determine the “go or no-go” decision and affect the design of phase III trials. We employed the ALC criterion to determine the sample sizes due to its practical advantage that the sample size can be calculated without specifying a loss or utility function. Moreover, the ALC criterion computes sample size subject to credible

interval, which is easy for clinicians who are less familiar with statistics to understand. Since it is not feasible to present a general closed form formula for ALC, we estimated the average length of posterior credible intervals and the sample size using the Monte Carlo method. Our simulation study showed that the Monte Carlo method performs well with relatively small values of S and M . Moreover, we developed a rapid grid search method to reduce the computational burden.

We focus on the natural conjugate prior family because many probability distributions from the exponential family, such as the Bernoulli, normal, Poisson, and exponential, are useful in the practice of phase II clinical trials. Another advantage is that the posterior distribution of conjugate models has a closed form density so we can easily derive quantities of interest analytically. In addition, the hyper-parameters of conjugate prior distributions are easy to understand because, in some sense, they can be interpreted as prior observations. However, one limitation is that some prior beliefs cannot be fully described by conjugate priors. In practice, prior elicitation should be carried out carefully to make sure that the incorporated information properly reflects experts' opinion and literature knowledge.

The main focus of this article is to introduce a design for randomized phase II clinical trials that computes a fixed pretrial sample size needed for providing reliable information to subsequent phase III trials. Our goal is to improve the completeness and reliability of data from phase II studies and ultimately increase the success rate of future phase III studies. Considering that the early stopping rules are sometimes important for clinical studies due to ethical or budget concerns, we also explore the applications of the proposed Bayesian approach in the adaptive settings. The idea is to first compute a pretrial sample size using the proposed design. Then after obtaining the first m observations, the stopping rules can be evaluated in terms of the posterior probability that μ_E exceeds μ_C based on the data accumulated so far. If the results of interim analysis suggest to proceed the trial, the sample size needed for the second part can be re-estimated using the updated posterior distributions as new priors.

We believe that the proposed phase II designs can improve current programs in many ways. The multiple constraints that protect the precision of information for treatment effect as well as for both new therapy and control arms will allow for more complete and reliable preliminary data for decision making. The extension to the entire exponential family will ensure the proposed design to be applicable to many important endpoints commonly used in phase II trials. The Bayesian framework equips phase II programs with enough flexibility to be applicable to any measures of central tendency and treatment effect. Finally, the usage of ALC criterion facilitates the sample size determination for phase II trials and makes the proposed design easily understandable by clinical experts.

Acknowledgements

We would like to thank the reviewers and the editor for their helpful and useful suggestions that improved the manuscript considerably.

References

- Adcock C (1988). A bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society. Series D (The Statistician)* 37, 433–439.
- Berry DA (1985). Interim analyses in clinical trials: classical vs. bayesian approaches. *Statistics in medicine* 4, 521–526. [PubMed: 4089353]
- Berry SM, Carlin BP, Lee JJ, and Muller P (2010). *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC press.
- Brunier HC and Whitehead J (1994). Sample sizes for phase ii clinical trials derived from bayesian decision theory. *Statistics in Medicine* 13, 2493–2502. [PubMed: 7701149]
- Ding M, Rosner GL, and Müller P (2008). Bayesian optimal design for phase ii screening trials. *Biometrics* 64, 886–894. [PubMed: 18162113]
- Eisenhauer E, Therasse P, Bogaerts J, Schwartz L, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer* 45, 228–247. [PubMed: 19097774]
- Food and Drug Administration (2010). Guidance for industry: adaptive design clinical trials for drugs and biologics. Retrieved from <https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf> (accessed August, 2014).
- Freedman LS and Spiegelhalter DJ (1989). Comparison of bayesian with group sequential methods for monitoring clinical trials. *Controlled clinical trials* 10, 357–367. [PubMed: 2691203]
- Gajewski BJ and Mayo MS (2006). Bayesian sample size calculations in phase ii clinical trials using a mixture of informative priors. *Statistics in medicine* 25, 2554–2566.
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB (2014). *Bayesian data analysis*. Boca Raton, FL: CRC press.
- Hand AL, Stamey JD, and Young DM (2011). Bayesian sample-size determination for two independent poisson rates. *Computer methods and programs in biomedicine* 104, 271–277. [PubMed: 21146897]
- Jiang W, Mahnken JD, He J, and Mayo MS (2016). Generalized optimal design for two-arm, randomized phase ii clinical trials with endpoints from the exponential dispersion family. *Pharmaceutical statistics* 15, 459–470. [PubMed: 27511063]
- Joseph L and Belisle P (1997). Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 209–226.
- Joseph L and Wolfson DB (1997). Interval-based versus decision theoretic criteria for the choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, 145–149.
- Joseph L, Wolfson DB, and Du Berger R (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society: Series D (The Statistician)* 44, 143–154.
- Karrison TG, Maitland ML, Stadler WM, and Ratain MJ (2007). Design of phase ii cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *Journal of the National Cancer Institute* 99, 1455–1461. [PubMed: 17895472]
- Lavin PT (1981). An alternative model for the evaluation of antitumor activity. *Cancer clinical trials* 4, 451–457. [PubMed: 7318127]
- Mandrekar SJ and Sargent DJ (2010). Randomized phase ii trials: time for a new era in clinical trial design. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer* 5, 932–934.
- Mayo MS and Gajewski BJ (2004). Bayesian sample size calculations in phase ii clinical trials using informative conjugate priors. *Controlled clinical trials* 25, 157–167. [PubMed: 15020034]
- Mayo MS, Mahnken JD, and Soong S. j. (2009). Optimal designs for two-arm, phase ii clinical trial design with multiple constraints. *Journal of biopharmaceutical statistics* 20, 106–124.
- Mehta CR and Cain K (1984). Charts for the early stopping of pilot studies. *Journal of Clinical Oncology* 2, 676–682. [PubMed: 6726306]

- Pezeshk H (2002). How many subjects?—a bayesian approach to the design of a clinical trial. *Iranian International Journal of Science* 3, 127–133.
- Piantadosi S (2005). *Clinical trials: a methodologic perspective*. Hoboken, NJ: John Wiley & Sons.
- Rubinstein L, Crowley J, Ivy P, LeBlanc M, and Sargent D (2009). Randomized phase ii designs. *Clinical Cancer Research* 15, 1883–1890. [PubMed: 19276275]
- Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, and Smith MA (2005). Design issues of randomized phase ii trials and a proposal for phase ii screening trials. *Journal of Clinical Oncology* 23, 7199–7206. [PubMed: 16192604]
- Stallard N (1998). Sample size determination for phase ii clinical trials based on bayesian decision theory. *Biometrics* 54, 279–294. [PubMed: 9544522]
- Stallard N, Thall PF, and Whitehead J (1999). Decision theoretic designs for phase ii clinical trials with multiple outcomes. *Biometrics* 55, 971–977. [PubMed: 11315037]
- Staquet M and Sylvester R (1977). A decision theory approach to phase ii clinical trials. *Biomedicine* 26, 262–266. [PubMed: 336106]
- Sylvester RJ (1988). A bayesian approach to the design of phase ii clinical trials. *Biometrics* 44, 823–836. [PubMed: 3203131]
- Tan S-B and Machin D (2002). Bayesian two-stage designs for phase ii clinical trials. *Statistics in medicine* 21, 1991–2012. [PubMed: 12111883]
- Taylor JM, Braun TM, and Li Z (2006). Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase ii design. *Clinical Trials* 3, 335–348. [PubMed: 17060208]
- Thall PF and Simon R (1994). Practical bayesian guidelines for phase iib clinical trials. *Biometrics* 50, 337–349. [PubMed: 7980801]
- Thall PF, Simon RM, and Estey EH (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in medicine* 14, 357–379. [PubMed: 7746977]
- Wieand HS (2005). Randomized phase ii trials: what does randomization gain? *Journal of clinical oncology* 23, 1794–1795. [PubMed: 15699476]
- Yin G (2013). *Clinical trial design: Bayesian and frequentist adaptive methods*, volume 876 Hoboken, New Jersey: John Wiley & Sons.
- Zhong W, Koopmeiners JS, and Carlin BP (2013). A two-stage bayesian design with sample size reestimation and subgroup analysis for phase ii binary response trials. *Contemporary clinical trials* 36, 587–596. [PubMed: 23583925]

Figure 1.

The Monte Carlo estimations of total sample size for 1 to 1 allocation ratio against $S = M$ for exponential distribution with Inverse-Gamma(3, 30) prior for control and Inverse-Gamma(3, 50) prior for experiment arms with $I_C = I_E = 11$ and $I_\delta = 0.9$. In each box plot, the central rectangle spans one standard deviation below the mean to one standard deviation above the mean. The horizontal line inside the rectangle shows the mean and the two “tails” above and below the box represent the maximum and minimum. The vertical dashed line corresponds to $M = S = 3000$. The point positions were jittered horizontally to reduce overlap.

Figure 2. Rapid grid search procedure. n_C^0 and n_E^0 represent the starting grid values that satisfy constraints (6) and (7). The arrows display the searching path.

Table 1

Bayesian optimal sample sizes for $I_C = I_E = 13:5$ and $I_\delta = 0:65$ under 1 to 1, 1 to 2, and optimal randomization allocation ratios for exponential distribution.

HR	M_C	M_E	Prior	1 to 1	1 to 2	Optimal	(n_C, n_E)
0.5	21.6	43.3	01	156	120	118	(40,78)
0.6	21.6	36.1	01	110	120	109	(59,50) (58,51) (57,52) (56,53) (55,54) (54,55) (53,56) (52,57) (51,58) (50,59)
0.5	21.6	43.3	02	230	174	158	(43, 115)
0.6	21.6	36.1	02	154	144	132	(55, 77)

HR stands for assumptions made for hazard ratio. The columns M_C and M_E represent the assumed modes for prior distributions of θ_C and θ_E . The column "Prior" corresponds to different prior distributions where "01" represents the informative prior (prior sample sizes of both arms are 20) and "02" represents the weakly informative prior (prior sample sizes of both arms are 10).