Published in final edited form as:

Contemp Clin Trials. 2013 November ; 36(2): . doi:10.1016/j.cct.2013.03.011.

A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials

Wei Zhong¹, Joseph S. Koopmeiners², and Bradley P. Carlin²

¹Department of Biostatistics, Genentech, Inc., 1 DNA Way, South San Francisco, CA, 94080

²Division of Biostatistics, MMC 303, University of Minnesota, Minneapolis, MN 55455

Abstract

Frequentist sample size determination for binary outcome data in a two-arm clinical trial requires initial guesses of the event probabilities for the two treatments. Misspecification of these event rates may lead to a poor estimate of the necessary sample size. In contrast, the Bayesian approach that considers the treatment effect to be random variable having some distribution may offer a better, more flexible approach. The Bayesian sample size proposed by Whitehead et al. (2008) for exploratory studies on efficacy justifies the acceptable minimum sample size by a “conclusiveness” condition. In this work, we introduce a new two-stage Bayesian design with sample size reestimation at the interim stage. Our design inherits the properties of good interpretation and easy implementation from Whitehead et al. (2008), generalizes their method to a two-sample setting, and uses a fully Bayesian predictive approach to reduce an overly large initial sample size when necessary. Moreover, our design can be extended to allow patient level covariates via logistic regression, now adjusting sample size within each subgroup based on interim analyses. We illustrate the benefits of our approach with a design in non-Hodgkin lymphoma with a simple binary covariate (patient gender), offering an initial step toward within-trial personalized medicine.

Keywords

Bayesian design; clinical trial; personalized medicine; predictive approach; sample size reestimation; subgroup analysis

1 Introduction

Traditional sample size determination for binary outcome data in a frequentist approach is simple, straightforward, and has been implemented in many clinical trials. For example, consider a two-arm clinical trial that compares the effect of two treatments, where we are interested in testing the hypotheses $H_0: p_1 = p_2$ versus $H_a: p_1 > p_2$, where p_1 and p_2 denote the true event rates in the two treatment groups. To obtain the sample size given some pre-specified significance level and power, we must first set some target point estimates of p_1 and p_2 as crude guesses of the event probabilities for two treatments, denoting them a p_1^*

© 2013 Elsevier Inc. All rights reserved.

Corresponding author: Bradley P. Carlin, phone: (612) 624-6646, fax: (612) 626-0660, brad@biostat.umn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

and p_2^* , respectively. The designed detectable effect is then $\theta^* = p_1^* - p_2^*$. The sample size can be calculated by the following standard formula (Lachin, 1977),

$$n \text{ per group} = \frac{2(Z_{1-\alpha/2} + Z_\beta)^2 \bar{p}^*(1 - \bar{p}^*)}{\theta^{*2}}, \quad (1)$$

where the average event rate $\bar{p}^* = (p_1^* + p_2^*)/2$, and Z denotes the percentile of the standard normal distribution.

Since the selection of the p_1^* and p_2^* are usually based on fairly vague prior knowledge or other studies with small sample sizes, the credibility of the “working alternative hypothesis” that $p_1 = p_1^*$ and $p_2 = p_2^*$ is often questionable (Spiegelhalter and Freedman, 1986). Misspecification of the event rates may lead to a poor estimate of the necessary sample size (Shih et al., 1997). To fix this problem, many sequential designs and adaptive sample size designs incorporating interim analyses have been proposed in recent years (Gehan, 1961; Simon, 1989; Jennison and Turnbull, 2000; Gould, 2001; Denne, 2001; Friede and Kieser, 2004). All these methods can provide substantial improvement by adjusting the sample size to achieve the target power while preserving the overall Type I error. However, previous sample size reestimation methods are based on an implicit assumption that estimates of the true unknown treatment effect do not change appreciably over time. In real life situations, this assumption is questionable, especially when more subject-level variability exists in the early recruitment period. A good specification of the expected treatment effect is still required for these frequentist designs.

In contrast, the Bayesian approach considers the treatment effect to be random variable having some distribution, and updates the prior with the data, obtaining a posterior distribution for inference. The interpretation of a credible interval for the treatment effect seems more natural here than that of the traditional frequentist confidence interval. Moreover, the objective of a phase II trial is to accept or reject a new drug for further investigation in a phase III trial, rather than obtain a highly precise estimate of each possible response rate. Generally there are three classes of Bayesian methods for sample size determination. First, a frequentist-Bayesian hybrid approach (Brown et al., 1987; Spiegelhalter et al., 1993; Lecoutre, 1999; Lee and Zelen, 2000), which considers the predictive probability of achieving the primary study goal based on the available data, but still aiming to control type I error. Second, some Bayesians recommend an interval length-based approach (Pham-Gia and Turkkan, 1992; Joseph et al., 1995; Pezeshk, 2003), which uses the length of posterior credible intervals as the sample size criterion. Finally, some authors pursue a fully decision-theoretic approach (Stallard, 1998; Claxton et al., 2000; Sahu and Smith, 2006; Berry et al., 2010), which chooses sample size to maximize an investigator-selected utility function or minimize a corresponding loss function.

The Bayesian sample size proposed by Whitehead et al. (2008) for exploratory studies on efficacy is an interval length-based approach, but includes an analogy to frequentist Type I and II errors. These authors argue that “the trial should be large enough to ensure that the data collected will provide convincing evidence either that an experimental treatment is better than a control or that it fails to improve upon control by some clinically relevant difference.” Like frequentist designs, the expected treatment effect is explicitly set in the design. But the Whitehead et al. sample size does not aim to meet certain power criteria under the alternative hypothesis. Instead, the acceptable minimum sample size N is justified by a “conclusiveness” condition. In the context of a one-sample test for a binary outcome (say, efficacy), it specifies that, regardless of the data, at least one of the two following probability statements should be satisfied at the end of a trial:

$$Pr(p > 0 | Y^N) \geq \eta_1 \text{ or } Pr(p < \theta^* | Y^N) \geq \eta_2, \quad (2)$$

where $p \in [0, 1]$ denotes the success rate for the treatment, $\theta^* \in [0, 1]$ is the expected (or desired) treatment effect, and Y^N represents any possible dataset of N patients. The threshold probabilities η_1 and η_2 are selected to reflect the degree of certainty we require for convincing evidence, with both values typically close to 1.

One potential problem is that such a sample size might be too conservative. Adding an interim stage to reestimate the sample size might offer a solution, dramatically reducing the sample size where the interim information about the true treatment effect emerges as sufficiently conclusive. Moreover, the corresponding Bayesian approach for comparing two proportions is not discussed by Whitehead et al. (2008) and merits further exploration.

At the interim stage, one can calculate the predictive power based on the interim posterior estimates of the parameters. The predictive power is actually the “re-estimated” power based on the prior and the data. Thus, a Bayesian approach to sample size estimation seems more sensible and natural here. However, in contrast to the frequentist literature, sample size reestimation has been infrequently discussed in the Bayesian setting. Some Bayesians argue that Bayesian analysis is a naturally sequential procedure, and are thus unconcerned about Type I error inflation resulting from multiple interim looks. Patient recruitment should depend on the data available at that time, and the adequacy of the resulting predictive power for making a final decision. However in practice, the sample size is usually determined before starting the trial and the schedule of interim analyses is also fixed; many trialists feel it is inappropriate to adjust the recruitment plan during the trial. Sample size reestimation, a key factor in interim analysis, is thus relevant in Bayesian design as well. Wang (2007) applies a Bayesian predictive approach to interim sample size reestimation, and compares it to other approaches such as predictive and conditional power approaches. The author recommends its application in exploratory studies, where knowledge about a test drug is still uncertain, and the adaptive sample size is based on the predictive probability of trial success.

“Personalized medicine” is a subject of intense discussion in recent years. The concept refers to the tailoring of treatments to individuals based on personal characteristics, and represents the next step in drug therapy and development toward better understanding of disease and health (Woodcock, 2007). The field is closely related to subgroup analysis, a subject of longstanding interest to trialists. For example, a recent study suggested no improvement in the overall mortality of patients with coronary disease whether treated with percutaneous coronary intervention (PCI) or coronary-artery bypass surgery. But the results also showed age played a key role, with much lower mortality after surgery among patients 65 years or older, while lower mortality after PCI among those 55 years or younger (Hlatky et al., 2009). Although many observational studies and pooled trials have contributed to our understanding of treatment effects at the individual level through subgroup effect analyses and development of prediction rules, a significant obstacle to the implementation of a personalized approach to trials themselves is the lack of appropriately designed studies (Garber and Tunis, 2009). Sample size estimation is an important issue for adequate trial design when we seek to study subgroup effects, especially in view of the well-known risk of Type I error inflation resulting from subgroups chosen post-hoc.

In this work, we first introduce a new two-stage Bayesian two-arm phase II trial design with sample size reestimation by implementing a predictive approach using the Whitehead et al. (2008) stopping rule. Then, we extend it to a four-subgroup trial design that considers an important binary covariate (gender) crossed with the treatment effect. Bayesian methods offer a direct attack on this problem, providing probabilities of efficacy, futility, and the like

given the data seen so far. Traditional frequentist tools do not do this, and in fact p -values tend to overstate the evidence against H_0 ; worse, they are often misinterpreted by non-statisticians as the probability that the null is true, even though this is far from accurate. While Bayesian methods have crucial design parameters that, like Type I and II error rates, determine the procedure's operating characteristics, they are not significantly more complex to implement in multi-center settings, yet capable of delivering results that are more easily interpreted by clinicians and patients and usually for a reduced total sample size.

The remainder of the article is organized as follows. In Section 2 we introduce our proposed two-stage Bayesian design in the case of a binary endpoint for a drug treatment trial. Section 3 presents the application of our design for a sample cancer trial with gender stratification, and compares its operating characteristics with those of the Whitehead et al. (2008) design. Finally, Section 4 discusses the advantages and limitations of our design, other applications, and suggests areas for further research.

2 Method

2.1 Initial Sample Size Calculation

In an equal-size two-arm phase II trial to test the efficacy difference between a drug and a placebo for a binary endpoint, we generalize (2) to provide an initial sample size N per group determined to satisfy at least one of the following two probability statements for any dataset of this size,

$$Pr(p_T - p_C > 0 | s_T, s_C) \geq \eta_1 \text{ or } Pr(p_T - p_C < \theta^* | s_T, s_C) \geq \eta_2, \quad (3)$$

where θ^* is the desired level of improvement in treatment response rate, p_T and p_C denote the success rates in the treatment and placebo groups, respectively, and s_T and s_C represent the numbers of efficacy events among N subjects in each group. Note that s_T and s_C can take any values in $\{0, 1, 2, \dots, N\}$. Suppose we place vague conjugate $Beta(\tau_T, T)$ and $Beta(\tau_C, C)$ priors on p_T and p_C . The posterior for p_T then follows a $Beta(\tau_T + s_T, T + N - s_T)$ distribution. In similar notation, the posterior for p_C follows a $Beta(\tau_C + s_C, C + N - s_C)$. Then, the exact posterior distribution of $(p_T - p_C)$ is denoted as the beta difference distribution, denoted by Pham-Gia and Turkkan (1993) as $BDI(\tau_T + s_T, T + N - s_T, \tau_C + s_C, C + N - s_C)$. Although the probability density function of this distribution is complex, Monte Carlo sampling from it is straightforward by generating from two independent beta distributions. In reality it is usually unacceptable to choose a sample size of less than 10 subjects per group, and thus we assume the smallest possible N to be 10. Thus, starting from $N = 10$, we check the criteria (3). If both $Pr(p_T - p_C > 0 | s_T, s_C) < \eta_1$ and $Pr(p_T - p_C < \theta^* | s_T, s_C) < \eta_2$ hold for any possible values of s_T and s_C , we increase N to $N + 1$ and continue to check until a minimum N is obtained that satisfies (3) for all s_T and s_C . Considering not much information will be available at the beginning of the trial, it is reasonable to begin with this rather conservative sample size. Note that this sample size only depends on the two threshold probabilities η_1 and η_2 , a selected prior on tumor response rates in each arm, and the target rate difference between two treatments we seek to detect.

2.2 Sample Size Reestimation

To add in a sample size reestimation step, we first assume all patients enroll simultaneously across groups, so that the numbers of patients at the interim look are the same for both groups. Suppose the efficacy data for N_1 patients are available for both groups at the interim look, with s_{T1} positive responses in the treatment group, and s_{C1} positive responses in the control group. N_1 could be determined as a proportion of the initial sample size; for example, $N_1 = \frac{1}{3}N$. Define the new sample size after recruiting N_1 patients to be M , and

assume there are s_{T2} and s_{C2} successful events among these M additional patients in the treatment and control groups, respectively. From the previous discussion, the interim posterior distribution of $(p_T - p_C)$ is then $BDF(\alpha_T + s_{T1} + s_{T2}, \alpha_T + N_1 + M - s_{T1} - s_{T2}; \alpha_C + s_{C1} + s_{C2}, \alpha_C + N_1 + M - s_{C1} - s_{C2})$, and can be easily sampled.

Our sample size reestimation is still based on a conclusiveness condition. Similar to (3), we are particularly interested in ensuring at least one of the following two conditions:

$$Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_1 \text{ or } Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_2. \quad (4)$$

Note that both M , s_{T2} and s_{C2} are unknown random variables at the interim stage, while N_1 , s_{T1} and s_{C1} are observed data. At this stage, since knowledge has been gained from the data of the first N_1 patients, it is reasonable to assume that the future data will be generated based on the interim posterior distributions of p_T and p_C . Unlike the consideration of all possible data for the initial sample size, this assumption may greatly decrease the potential expected variability of the future data, thus substantially decreasing the sample size.

Given M , s_{T2} and s_{C2} can be predicted using beta-binomial marginal distributions having pmfs,

$$P(s_{T2}) = \frac{B(\alpha_T + s_{T1} + s_{T2}, \beta_T + N_1 + M - s_{T1} - s_{T2})}{(M+1)B(M - s_{T2} + 1, s_{T2} + 1)B(\alpha_T + s_{T1}, \beta_T + N_1 - s_{T1})}, \quad (5)$$

and

$$P(s_{C2}) = \frac{B(\alpha_C + s_{C1} + s_{C2}, \beta_C + N_1 + M - s_{C1} - s_{C2})}{(M+1)B(M - s_{C2} + 1, s_{C2} + 1)B(\alpha_C + s_{C1}, \beta_C + N_1 - s_{C1})}, \quad (6)$$

where B is the beta function, $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, and $s_{T2}, s_{C2} \in \{0, \dots, M\}$. Following Sec 4.2 of Berry et al. (2010), the predictive probabilities we require can be obtained as

$$Pr_{pred1} = \sum_{s_{T2}=0}^M \sum_{s_{C2}=0}^M I\{Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_1\} P(s_{T2}) P(s_{C2}), \quad (7)$$

and

$$Pr_{pred2} = \sum_{s_{T2}=0}^M \sum_{s_{C2}=0}^M I\{Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \geq \eta_2\} P(s_{T2}) P(s_{C2}), \quad (8)$$

where $I(\cdot)$ is the indicator function, taking the value 0 or 1, depending on whether the condition is false or true. Predictive probabilities are a standard tool in Bayesian clinical trial analysis, used anytime we need to make a decision in the face of uncertainty about yet-to-be-observed data (in our case, the Stage 2 data). We select the minimum sample size M so that at least one of the two predictive probabilities is no less than a desired level α . A special case is that M could be 0, in which case $Pr(p_T - p_C > 0 | N_1, s_{T1}, s_{C1}) \geq \eta_1$ or $Pr(p_T - p_C < \theta^* | N_1, s_{T1}, s_{C1}) \geq \eta_2$. This situation suggests the interim data are already strong enough to reach a decision to stop the trial at that point due to either efficacy or futility, respectively.

When M is large, the computation in (7) and (8) is extensive. The sampling time to get the BDI distribution of $p_T - p_C$ and the calculation of $P(s_{T2})$ and $P(s_{C2})$ for all possible values of s_{T1} and s_{C1} are the two main factors. The use of the normal approximation arising from the so-called Bayesian Central Limit Theorem (BCLT) can speed the computation; see Section 3.2 of Carlin and Louis (2009) for details. As $(N_1 + M)$, the posterior distribution of $(p_T - p_C)$ at the interim stage can be approximated by a Normal distribution,

$$(p_T - p_C | N_1, s_{T1}, s_{C1}, M, s_{T2}, s_{C2}) \rightsquigarrow N \left(\hat{p}_T - \hat{p}_C, \frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{N_1 + M} \right), \quad (9)$$

where $\hat{p}_T = \frac{s_{T1} + s_{T2}}{N_1 + M}$ and $\hat{p}_C = \frac{s_{C1} + s_{C2}}{N_1 + M}$ are the MLE estimators of p_T and p_C in a frequentist version. Thus (7) and (8) can be approximated as

$$Pr_{pred1} \approx \sum_{s_{T2}=0}^M \sum_{s_{C2}=0}^M I \left\{ \Phi \left(\frac{\hat{p}_T - \hat{p}_C}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{N_1 + M}}} \right) \geq \eta_1 \right\} P(s_{T2}) P(s_{C2}), \quad (10)$$

and

$$Pr_{pred2} \approx \sum_{s_{T2}=0}^M \sum_{s_{C2}=0}^M I \left\{ \Phi \left(\frac{\theta^* - (\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{N_1 + M}}} \right) \geq \eta_2 \right\} P(s_{T2}) P(s_{C2}), \quad (11)$$

where $\Phi(x)$ represents the cumulative density function of a standard normal distribution. Alternatively, to also avoid calculating $P(s_{T2})$ and $P(s_{C2})$ in (10) and (11) for all possible values of s_{T1} and s_{C1} , we can approximate (10) and (11) by directly sampling s_{T2} and s_{C2} from their beta-binomial distributions (5) and (6), obtaining $\{(s_{T2j}^*, s_{C2j}^*), j = 1, \dots, J\}$. Then (10) and (11) can be approximated as

$$Pr_{pred1} \approx \frac{1}{J} \sum_{j=1}^J I \left\{ \Phi \left(\frac{\hat{p}_{Tj}^* - \hat{p}_{Cj}^*}{\sqrt{\frac{\hat{p}_{Tj}^*(1-\hat{p}_{Tj}^*) + \hat{p}_{Cj}^*(1-\hat{p}_{Cj}^*)}{N_1 + M}}} \right) \geq \eta_1 \right\}, \quad (12)$$

and

$$Pr_{pred2} \approx \frac{1}{J} \sum_{j=1}^J I \left\{ \Phi \left(\frac{\theta^* - (\hat{p}_{Tj}^* - \hat{p}_{Cj}^*)}{\sqrt{\frac{\hat{p}_{Tj}^*(1-\hat{p}_{Tj}^*) + \hat{p}_{Cj}^*(1-\hat{p}_{Cj}^*)}{N_1 + M}}} \right) \geq \eta_2 \right\}, \quad (13)$$

where $\hat{p}_{Tj}^* = \frac{s_{T1} + s_{T2j}^*}{N_1 + M}$ and $\hat{p}_{Cj}^* = \frac{s_{C1} + s_{C2j}^*}{N_1 + M}$. See Section 3.3 of Carlin and Louis (2009) for a review of noniterative Monte Carlo methods. This sampling approach avoids extensive computation on $(M+1)^2$ different (s_{T2}, s_{C2}) pairs, potentially valuable when M is large.

2.3 Final Trial Conclusion

At the end of the trial, if the final results show $Pr(p_T - p_C > 0 | \text{all data}) \geq 1 - \alpha$ or $Pr(p_T - p_C < -\alpha | \text{all data}) \geq \alpha$, then we conclude that the test drug is minimally effective, or not as effective as we had hoped, respectively. Note in particular that our Whitehead formulation means that both conclusions may be drawn simultaneously. This is clearly an odd situation, but one that is inherent in Whitehead et al.'s definition of "conclusiveness," arising when the posterior mass piles up in the "indifference zone" $(-\alpha, \alpha)$. This is likely whenever any of the true treatment effects are near $\alpha/2$, the midpoint of this interval. The implications of this case are that the drug effect is likely to be only around half of what had been hoped, and thus we may need to think harder about the clinical significance of our original choice of α . Alternatively, despite our best efforts, it may be that *neither* of the previous two probability statements holds, suggesting either bad luck or perhaps a data pattern for later patients that is different from that of the interim data. For temporally homogeneous data, this situation should happen rarely provided we choose the predictive probability of conclusiveness reasonably close to 1. Generally our proposed two-stage Bayesian sample size algorithm for each stratum is as follows:

Sample Size Estimation Algorithm

1. Calculate the initial sample size based on the pre-specified expected effect.
2. Choose a proportion of the initial sample size, N_1 , to determine the interim time for sample size reestimation.
3. Begin the trial, and estimate the interim posterior distribution of the treatment effect at the interim time (i.e., when N_1 patients have reported their data in each treatment group). Check the conclusiveness condition, thus making a decision on whether to stop the trial there.
4. If not terminated, find the minimum sample size M so that at least one of the two predictive probabilities (7) and (8) is over a desired level γ ; this is our reestimated Bayesian sample size.
5. Resume the trial, and make a conclusion on the efficacy of the drug at the end of the trial, after all $N_1 + M$ patients in each group have reported.

2.4 Extension to Personalized Medicine

As mentioned in Section 1, patient demographics such as gender and age may be highly correlated with treatment effect. Thus, these confounding factors should be considered when testing for a treatment effect, and constitute a first step toward within-trial consideration of personalized medicine. Without loss of generality, let $Y_i \in \{0, 1\}$ be the tumor response for the i th patient ($i = 1, \dots, n$), and assume the Y_i follow a Bernoulli distribution with success rate p_i . Then we apply a logistic regression model for p_i as follows,

$$p_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})}, \quad (14)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ is a vector of covariate values for the i th patient, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients ($k \geq 3$). We set $x_{i1} = 1$ for the estimation of the intercept term, and let $x_{i2} = 1$ or 0 for the treatment and placebo groups, respectively. Interaction terms, such as the multiplicative interaction between drug and gender, may also be desired in $\boldsymbol{\beta}$. Suppose there are L different $(x_{i3}, \dots, x_{ik})$ covariate vector values in the study cohort, and all patients can be classified into one of these L strata, with a particular $(x_{i3}, \dots, x_{ik})$ for each stratum. We also assume equal sample size for both the drug and placebo arms within

each stratum. If continuous covariates such as age and weight are included in the logistic model, then we can create a set of ranges based on their values, and use dummy variables to control the number of strata L .

Our two-stage Bayesian design can be applied to the multiple strata case, where we now aim to make a conclusiveness statement on the drug effect in each stratum. When L is small, it is easy to directly use beta priors for the probabilities of success events in each subgroup to adjust sample size within each stratum. We illustrate such subgroup analysis with a numerical example in Section 3. The sample sizes for both treatment groups are assumed equal, but we permit different sample sizes between groups with different covariate values. The four-subgroup trial includes two sub-trials proposed by our method for the two covariate groups, with final conclusions made for both groups; we simply apply the algorithm in the previous subsection to every stratum.

However, when L is large, priors are often set for β in the logistic model, rather than directly for the stratum-specific p_j in (14). Note that it is not important here to find the initial sample size so as to meet the conclusiveness condition regardless of the data, since logistic models are usually less than ideal in their handling of extreme data (i.e., when the number of success events is low even though the total number of patients is large). Our method enables the borrowing of strength from the interim data in sample size reestimation to make a conclusion on drug treatment effects.

Under either a uniform or a normal prior specification, the posterior distribution for β is complicated and does not have an analytic closed form. Although MCMC software may be used to obtain the exact distribution, the computation is extensive if using the predictive approach to test all possible datasets with the new sample size. Therefore, we again suggest use of the BCLT to simplify computation.

Suppose a uniform (improper) prior is specified for β , i.e., $P(\beta) \propto 1$. Then the posterior distribution of β follows a multivariate normal distribution as $n \rightarrow \infty$,

$$\beta|Y \rightsquigarrow MVN_k(\hat{\beta}, (X' \hat{V} X)^{-1}), \quad \text{as } n \rightarrow \infty, \quad (15)$$

where $Y = (Y_1, Y_2, \dots, Y_n)$ is the binary outcome data, $\hat{\beta}$ is the MLE of β , and $X = (X_1, X_2, \dots, X_n)$ is an $n \times k$ covariate matrix. Then V is an $n \times n$ diagonal matrix with i th diagonal element:

$$V_{ii} = \frac{\exp(\mathbf{x}_i' \hat{\beta})}{\{1 + \exp(\mathbf{x}_i' \hat{\beta})\}^2}.$$

Suppose the i th and j th patients are in the drug and placebo arms within the same l th stratum, respectively ($1 \leq l \leq L$). The difference in the probabilities of tumor response between the new drug and placebo within this stratum is

$$p_i - p_j = \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} - \frac{\exp(\mathbf{x}_j^T \beta)}{1 + \exp(\mathbf{x}_j^T \beta)}. \quad (16)$$

Let $\mathbf{x}_i = (1, 1, \dots)$ and $\mathbf{x}_j = (1, 0, \dots)$ be the covariate vectors for the drug and placebo arms, respectively. By the delta method, the posterior distribution of $p_i - p_j$ can be approximated by a normal distribution,

$$\begin{aligned}
 p_i - p_j | \text{data} &\rightsquigarrow N \left(\frac{\exp(\mathbf{x}'_i \beta)}{1 + \exp(\mathbf{x}'_i \beta)} - \frac{\exp(\mathbf{x}'_j \beta)}{1 + \exp(\mathbf{x}'_j \beta)}, (X_{i\beta} + X_{j\beta})' (X' \hat{V} X)^{-1} (X_{i\beta} + X_{j\beta}) \right), \\
 \text{where } X_{i\beta} &= \left(\frac{\exp(\mathbf{x}'_i \hat{\beta})}{\{1 + \exp(\mathbf{x}'_i \hat{\beta})\}^2} x_{i1}, \frac{\exp(\mathbf{x}'_i \hat{\beta})}{\{1 + \exp(\mathbf{x}'_i \hat{\beta})\}^2} x_{i2}, \dots, \frac{\exp(\mathbf{x}'_i \hat{\beta})}{\{1 + \exp(\mathbf{x}'_i \hat{\beta})\}^2} x_{ik} \right)', \\
 \text{and } X_{j\beta} &= \left(\frac{\exp(\mathbf{x}'_j \hat{\beta})}{\{1 + \exp(\mathbf{x}'_j \hat{\beta})\}^2} x_{j1}, \frac{\exp(\mathbf{x}'_j \hat{\beta})}{\{1 + \exp(\mathbf{x}'_j \hat{\beta})\}^2} x_{j2}, \dots, \frac{\exp(\mathbf{x}'_j \hat{\beta})}{\{1 + \exp(\mathbf{x}'_j \hat{\beta})\}^2} x_{jk} \right)',
 \end{aligned} \quad (17)$$

Similarly, we can check the probability criteria as in (7) and (8) in a predictive approach to reestimate sample size at the interim stage. In the presence of covariates leading to L strata, we can design a multi-subgroup trial by combining multiple sub-trials as proposed above. Note that we would likely power the study at the interim look for each stratum separately; Section 3 offers an illustration. Alternatively, unpromising strata might be abandoned at the interim look as a cost-saving measure. Also, note our current formulation defines conclusiveness within each subgroup, so that, for example, a trial could be conclusive for men but not for women. While we do not add an “overall conclusiveness” condition, such a condition could certainly be formulated and added if desired.

3 Application

In this section, we give a simple but representative example, which concerns a design of phase II trial for patients with non-Hodgkin lymphoma. The primary goal of this study is to design a trial with an efficient sample size to assess the efficacy of a novel natural killer (NK) cell treatment compared to placebo. A decision as to whether the new regimen deserves a test in a large confirmatory phase III trial must be made at the conclusion of this study. The decrease in the expression level for polychlorinated biphenyl (PCB) is considered as the tumor response. Little information is available for both treatments at the planning stage, but it is expected that the treatment effects may be different in women and men. Therefore, a two-arm randomized trial stratified by gender is preferred. As in the previous section, this four-subgroup trial design reestimates the sample size after one interim analysis, to improve the chance of a conclusive decision regarding the treatment effect for both women and men. We also permit different sample sizes in the two groups.

In this study, the target difference in tumor response rates between the two treatments is set as $\delta^* = 0.2$. To retain the exploratory nature of our design, the threshold probabilities α_1 and α_2 are both fixed at 90%, fairly modest values. The proportion of the initial sample size for the interim time is fixed at $N_1 = N/3$. The desired conclusiveness level γ , or the threshold of predictive probability of conclusiveness, is set as 0.9. We apply very weak independent Beta(0.5,0.5) priors to the tumor response rates in all four subgroups: Trt+Women, Control+Women, Trt+Men and Control+Men.

We assume the binary outcome data is independently generated from binomial distributions with different tumor response rates in all four subgroups. To evaluate our new design, we consider the six different simulation scenarios given in Table 1. The true tumor response rates for all four subgroups in each scenario are listed as p_{TW} , p_{CW} , p_{TM} and p_{CM} . The p_W and p_M represent the true effect differences between the treatments for women and men, respectively. The first scenario assumes that the new drug has a better treatment effect than placebo in both women and men (0.3 and 0.2, respectively). Scenario 2 instead supposes that the new treatment is superior to placebo only in women, but not in men (the differences are 0.2 and 0, respectively). In Scenario 3, no treatment difference exists between the two treatments across both gender strata. Scenario 4 indicates that the new drug is marginally better than placebo in both women and men, with the improvement in success rate as 0.1,

less than the hoped-for value of $\delta^* = 0.2$. The next scenario shows the improvement in men is 0.3, while that in women is only 0.1, marginally better. The last scenario assumes that the new treatment is much more effective than placebo in women, but even *worse* than placebo in men.

First, we calculate the initial sample size for each subgroup as in Subsection 2.1. The minimum sample size that is sufficient to make a conclusive statement on the treatment effect is 80 in each of the four subgroups. At the interim stage ($N_1 = 80 \times 1/3 \approx 27$), our design reestimates the sample size using the interim data by the predictive probability approach described in Section 2. Then, the conclusiveness condition is examined predictively using the new sample size. We simulated 1000 trials to investigate the operating characteristics of this approach using the R software. As a comparison, the performance of simply using the initial sample size generalized from Whitehead et al. (2008) without any sample size reestimation is also investigated.

Table 2 presents the simulated operating characteristics of our design, including the average sample size per subgroup, probabilities of early termination (PET) at the interim time, and conclusiveness probabilities at the end of the trial for women and men, respectively. Operating characteristics for the fixed-sample design using the generalized Whitehead et al. sample size can be found in Table 3. Both tables further break down the conclusiveness probabilities into their “minimal efficacy” and “futility” components.

Comparing the results in Tables 2 and 3, our proposed design results in a substantial decrease in sample size for all scenarios. Whereas the fixed-sample design requires 80 men and 80 women, our design requires an average of 28–45 women and 32–44 men, depending on the scenario. Furthermore, we achieve conclusiveness in all cases, even though we require only half the sample size on average.

One drawback to our proposed design is that we observe an increased probability of reaching an incorrect conclusion. For example, in Scenario 1 where $p_M = 0.2$, we conclude that $p_M < 0.2$ about 18% of time, compared to only 11% for the fixed-sample design. A similar phenomenon is observed in Scenarios 2 and 3. These results should not be surprising. Although our model is estimated using the Bayesian paradigm, the results reported in Table 2 represent the frequentist operating characteristics of our proposed design. As in any sequential procedure, allowing early termination at the interim look increases the Type I and Type II error rates.

One approach to improving the percentage of correct conclusions is to adopt stricter stopping criteria. For example, we may raise the probability thresholds α_1 and α_2 from 0.9 to 0.95, which also increases the maximum sample size from 80 to 131 per subgroup for both women and men. The operating characteristics of our design under this setting are shown in Table 4. This change decreases the probability of reaching an incorrect conclusion to that observed for the fixed-sample case (Table 3). Increasing α_1 and α_2 also increases the expected sample size compared to Table 2, but the expected sample sizes are still less than those in the fixed-sample case in all but the most difficult scenarios, namely when p_M or $p_W = 0.1$ (exactly halfway between 0 and δ^*).

For comparison, we also consider the operating characteristics of a standard frequentist group sequential design. There is no true frequentist analog to the conclusiveness condition proposed by Whitehead, but a reasonable comparison is to a design with type I error equal to its type II error. This assures that the $(1 - \alpha)$ -level confidence interval would not cover both the null and alternative hypotheses at study completion, and is similar in spirit to our conclusiveness condition. From Table 4, we see that our design has approximate type I error equal to 0.07 and power approximately equal to 0.93 for detecting a difference of 0.20.

Therefore, we will compare the operating characteristics of our study to a frequentist group sequential design with a single interim analysis and stopping boundaries derived using the method proposed by Emerson and Fleming (1989). Table 5 presents the operating characteristics of this frequentist design with a single interim analysis after one third of the subjects have been enrolled in the trial. Operating characteristics for this design were calculated using the RCTdesign package for R; see <http://www.rctdesign.org/Welcome.html>. This design has a smaller maximum sample size than our design (102 subjects/group, compared to 131 subjects/group for our design), but we see that the Bayesian design yields a smaller *average* sample size for both genders across all scenarios while maintaining operating characteristics similar to those of the frequentist design.

4 Discussion

Our design inherits the properties of good interpretation and easy implementation from Whitehead et al. (2008), generalizes their method to a two-sample setting, and uses a fully Bayesian predictive approach to reduce an unnecessarily large sample size and save patients in exploratory studies. Moreover, we extend our method to multiple subgroups with varying categorical covariates, and allow flexible sample size within each subgroup based on interim analyses. With these merits, our design might be applied to many early phase studies, with consequent advances for personalized medicine.

Due to the different context between frequentists and Bayesians, when evaluating operating characteristics of a Bayesian approach by simulated data from fixed treatment effects, there is always an issue of how to best select control parameters such as α_1 , α_2 and β . Our results suggest that using stricter criteria in our method can reduce some errors and improve operating characteristics while not increasing sample size beyond Whitehead et al. levels. We also tried increasing β from 0.9 to 0.95, but the corresponding impacts on the sample sizes and the percentages of making correct conclusions were not as great.

Our formulation essentially assumes no population shift between Stages 1 and 2. Such a shift over time would seem unlikely provided the methods for selecting the patients did not change dramatically over time. Moreover, as long as patients are randomized to treatment assignment, the effect of any such shift on the treatment effect should be minimal. Finally, such a shift could only be estimated at the very end of the trial, since at the interim look when the sample size reestimation takes place, no Stage 2 data would yet be available. Still, recent work by Hobbs et al. (2011) could be used to address the possibility of a shift over time by making a decision regarding the “commensurability” of the data at Stages 1 and 2.

A referee has asked about the size of J needed to make (12) and (13) good Monte Carlo (MC) approximations to (10) and (11), respectively. Table 6 offers a comparison of the $P_{r_{pred1}}$ and $P_{r_{pred2}}$ estimates obtained via the two approaches, using $M = 100$ and $J = 5,000$ and 10,000. Since the predictive probabilities depend crucially on the values of N_1 , s_{T1} , and s_{T2} , we investigated eight different cases, four using $N_1 = 50$ and four using $N_1 = 100$. Results indicate that the MC approximation is quite good for both values of J , with absolute biases (compared to the non-MC version) less than 0.01 in all cases. Still, the largest biases (those exceeding 0.005) all arise using $J = 5,000$, so we recommend using $J = 10,000$ to obtain good results in practice.

Of course, there are also some limitations to our method. First, although we use logistic models for the case of many strata, the probability criteria are hard to determine because the interpretation for the transformation from the odds ratio to the absolute probability difference is not straightforward and depends on the baseline probability. Still more assumptions are required for the use of logistic regression models in clinical trials. For example, logistic regression often ignores some interaction terms between different

covariates to decrease the number of model coefficients. In addition, in this paper we assumed independent binomial models for all subgroups, without considering any correlations between them. How to incorporate correlations into our design is a topic for future study, as is adding response-adaptive randomization into our design; see e.g. Berry et al. (2010, Sec. 4.4).

Finally, although the selection of the interim time and how it influences operating characteristics can be investigated by simulation studies, our future work also includes finding more general guidance for choosing N_1 , the interim sample size, a subject that is also often discussed in frequentist sequential analysis. In our setting, we experimented with $N_1 = N/2$ and $N/6$, two values on either side of our previously chosen value of $N_1 = N/3$. Since the average total sample size $E(N) = N_1 + E(M)$, the selection of N_1 can affect $E(N)$ in potentially different ways. If the underlying truth is consistent with strongly conclusive evidence, then a smaller N_1 may still lead to a high probability of early termination ($M=0$), and thus a smaller overall sample size. For instance, in Scenario 6 for women, lowering N_1 to 14 (roughly $N/6$) from 27 lowers $E(N_W)$ in Table 2 from 28.17 to 25.15; PET_W remains fairly high at 0.83. However, in more equivocal settings, smaller N_1 may also yield substantially lower probabilities of early termination, potentially increasing $E(N)$. Indeed, in Scenario 6 for men, the aforementioned lowering of N_1 to 14 leads to a sharp drop in PET_M to 0.68, and a consequent rise in $E(N_M)$ to 34.99, an increase of about 2.5 patients. Although it is hard to pick a single interim sample size that is just right for all subgroups without seeing any data, comparison of our Table 2 results to those using smaller and larger values of N_1 suggests that our original value of $N_1 = N/3$ offers a good middle ground, producing lower $E(N_W)$ and $E(N_M)$ values in most of the six scenarios.

References

- Berry, S.; Carlin, B.; Lee, J.; Muller, P. Bayesian Adaptive Methods for Clinical Trials. CRC Press; 2010.
- Brown B, Herson J, Neely Atkinson E, Elizabeth Rozell M. Projection from previous studies: a bayesian and frequentist compromise. Controlled Clinical Trials. 1987; 8:29–44. [PubMed: 3568694]
- Carlin, B.; Louis, T. Bayesian methods for data analysis. Vol. 78. Chapman & Hall/CRC; 2009.
- Claxton K, Lacey L, Walker S. Selecting treatments: a decision theoretic approach. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2000; 163:211–225.
- Denne J. Sample size recalculation using conditional power. Statistics in Medicine. 2001; 20:2645–2660. [PubMed: 11523074]
- Emerson S, Fleming T. Symmetric group sequential test designs. Biometrics. 1989; 45:905–923. [PubMed: 2675998]
- Friede T, Kieser M. Sample size recalculation for binary data in internal pilot study designs. Pharmaceutical Statistics. 2004; 3:269–279.
- Garber A, Tunis S. Does comparative-effectiveness research threaten personalized medicine? New England Journal of Medicine. 2009; 360:1925–1927. [PubMed: 19420360]
- Gehan E. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. Journal of Chronic Diseases. 1961; 13:346–353. [PubMed: 13704181]
- Gould A. Sample size re-estimation: recent developments and practical considerations. Statistics in Medicine. 2001; 20:2625–2643. [PubMed: 11523073]
- Hlatky M, Boothroyd D, Bravata D, Boersma E, Booth J, Brooks M, Carrié D, Clayton T, Danchin N, Flather M, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. The Lancet. 2009; 373:1190–1197.

- Jennison, C.; Turnbull, B. Group Sequential Methods with Applications to Clinical Trials. CRC Press; 2000.
- Joseph L, Wolfson D, Du Berger R. Some comments on Bayesian sample size determination. *The Statistician*. 1995; 44:167–171.
- Lachin J. Sample size determinations for $r \times c$ comparative trials. *Biometrics*. 1977; 33:315–324. [PubMed: 884194]
- Lecoutre B. Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*. 1999; 79:93–105.
- Lee S, Zelen M. Clinical trials and sample size considerations: another perspective. *Statistical Science*. 2000; 15:95–110.
- Pezeshk H. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research*. 2003; 12:489–504. [PubMed: 14653352]
- Pham-Gia T, Turkkan N. Sample size determination in Bayesian analysis. *The Statistician*. 1992; 41:389–397.
- Pham-Gia T, Turkkan N. Bayesian analysis of the difference of two proportions. *Communications in Statistics-Theory and Methods*. 1993; 22:1755–1771.
- Sahu S, Smith T. A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2006; 169:235–253.
- Shih W, Zhao P, et al. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine*. 1997; 16:1913–1923. [PubMed: 9304763]
- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
- Spiegelhalter D, Freedman L. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*. 1986; 5:1–13. [PubMed: 3961311]
- Spiegelhalter D, Freedman L, Parmar M. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine*. 1993; 12:1501–1511. [PubMed: 8248659]
- Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics*. 1998; 54:279–294. [PubMed: 9544522]
- Wang M. Sample size reestimation by Bayesian prediction. *Biometrical Journal*. 2007; 49:365–377. [PubMed: 17623342]
- Whitehead J, Valdés-Márquez E, Johnson P, Graham G. Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*. 2008; 27:2307–2327. [PubMed: 18069728]
- Woodcock J. The prospects for “personalized medicine” in drug development and drug therapy. *Clinical Pharmacology & Therapeutics*. 2007; 81:164–169. [PubMed: 17259943]

The numeric settings for all four subgroups in six different scenarios. The p_{TW} represents the true tumor response rate in women with the new drug treatment. The p_{CW} is for women with placebo treatment, p_{TM} is for men with new drug treatment, and p_{CM} is for men with placebo treatment. The p_W and p_M represent the treatment differences in women and men, respectively.

Table 1

Scenario	p_{TW}	p_{CW}	p_{TM}	p_{CM}	p_W	p_M
1	0.5	0.2	0.5	0.3	0.3	0.2
2	0.4	0.2	0.3	0.3	0.2	0
3	0.3	0.3	0.4	0.4	0	0
4	0.4	0.3	0.3	0.2	0.1	0.1
5	0.4	0.3	0.5	0.2	0.1	0.3
6	0.6	0.2	0.3	0.4	0.4	-0.1

Table 2

Operating characteristics of our design. The $E(N_W)$ and $E(N_M)$ denote the average numbers in each treatment group for women and men, respectively. The PET_W and PET_M represent the probabilities of early termination due to conclusiveness at the interim time for women and men, respectively ($N_{Wmax} = N_{Mmax} = 80, \alpha = 0.05, \beta = 0.1, \gamma = 0.9, \delta = 0.9$).

Scenario	Women					Men				
	$E(N_W)$	PE_{TW}	$P(\text{efficacy}) (p_W > 0)$	$P(\text{futility}) (p_W < 0.2)$	$P(\text{conclusive}) (W \text{ overall})$	$E(N_M)$	PE_{TM}	$P(\text{efficacy}) (p_M > 0)$	$P(\text{futility}) (p_M < 0.2)$	$P(\text{conclusive}) (M \text{ overall})$
1	32.88	0.89	0.97	0.03	1.00	42.53	0.71	0.83	0.18	1.00
2	41.15	0.73	0.86	0.16	1.00	40.09	0.75	0.15	0.87	1.00
3	40.67	0.74	0.15	0.86	1.00	43.80	0.68	0.16	0.85	1.00
4	45.44	0.65	0.49	0.53	1.00	43.32	0.69	0.49	0.56	1.00
5	45.39	0.65	0.52	0.50	1.00	33.15	0.88	0.98	0.02	1.00
6	28.17	0.98	1.00	0.00	1.00	32.46	0.90	0.02	0.98	1.00

Table 3
Operating characteristics of the design generalized from Whitehead et al. (2008) without any sample size reestimation ($N_W = N_M = 80$).

Scenario	Women			Men		
	P(efficacy) ($p_W > 0$)	P(futility) ($p_W < 0.2$)	P(conclusive) (W overall)	P(efficacy) ($p_M > 0$)	P(futility) ($p_M < 0.2$)	P(conclusive) (M overall)
1	1.00	0.00	1.00	0.91	0.11	1.00
2	0.92	0.12	1.00	0.11	0.92	1.00
3	0.09	0.94	1.00	0.12	0.90	1.00
4	0.51	0.56	1.00	0.58	0.59	1.00
5	0.53	0.53	1.00	1.00	0.00	1.00
6	1.00	0.00	1.00	0.00	1.00	1.00

Table 4

Operating characteristics of our design ($N_{Wmax} = N_{Mmax} = 131$, $\alpha = 0.05$, $\beta = 0.9$).

Scenario	Women					Men				
	$E(N_W)$	PET_W	$P(\text{efficacy}) (p_W > 0)$	$P(\text{futility}) (p_W < 0.2)$	$P(\text{conclusive}) (W \text{ overall})$	$E(N_M)$	PET_M	$P(\text{efficacy}) (p_M > 0)$	$P(\text{futility}) (p_M < 0.2)$	$P(\text{conclusive}) (M \text{ overall})$
1	51.48	0.91	1.00	0.00	1.00	72.45	0.67	0.92	0.08	1.00
2	68.10	0.72	0.93	0.08	1.00	65.49	0.75	0.07	0.94	1.00
3	66.53	0.74	0.06	0.95	1.00	71.75	0.68	0.07	0.94	1.00
4	84.72	0.53	0.52	0.52	1.00	80.02	0.59	0.55	0.55	1.00
5	86.80	0.51	0.52	0.52	1.00	51.13	0.92	0.99	0.01	1.00
6	44.87	0.99	1.00	0.00	1.00	50.00	0.93	0.01	0.99	1.00

Operating characteristics for a frequentist design with type I and type II errors equal to 0.07, and a single interim analysis after one third of the subjects have been observed.

Table 5

Scenario	Women				Men			
	$E(N_w)$	PET_w	$P(\text{reject})$	$P(\text{fail to reject})$	$E(N_m)$	PET_m	$P(\text{reject})$	$P(\text{fail to reject})$
1	67.74	0.50	1.00	0.00	88.06	0.20	0.93	0.07
2	88.06	0.20	0.93	0.07	88.06	0.20	0.07	0.93
3	88.06	0.20	0.07	0.93	88.06	0.20	0.07	0.93
4	95.81	0.09	0.50	0.50	95.81	0.09	0.50	0.50
5	95.81	0.09	0.50	0.50	67.74	0.50	1.00	0.00
6	47.12	0.80	1.00	0.00	67.74	0.50	0.00	1.00

Table 6

Investigation of the quality of the Monte Carlo (MC) approximation in (12) and (13) using two values of J when $M = 100$.

Case	N_1	s_{T1}	s_{C1}	Non-MC				MC ($J = 10,000$)				MC ($J = 5,000$)			
				P	r_{pred1}	P	r_{pred2}	P	r_{pred1}	P	r_{pred2}	P	r_{pred1}	P	r_{pred2}
1	50	18	12	0.7565	0.1215	0.7538	0.1266	0.7660	0.1152						
2	50	15	10	0.6939	0.1860	0.6923	0.1828	0.7022	0.1798						
3	50	30	10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000						
4	50	15	5	0.9862	0.0069	0.9848	0.0078	0.9834	0.0079						
5	100	18	12	0.7565	0.1215	0.7598	0.1186	0.7599	0.1180						
6	100	15	10	0.6939	0.1860	0.6851	0.1830	0.7024	0.1804						
7	100	30	10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000						
8	100	15	5	0.9862	0.0069	0.9853	0.0072	0.9874	0.0058						