Predictive Posterior Power for Sample Size Re-estimation

Marc Sobel and Ibrahim Turkoz

Dept of Statistics, Temple University

Janssen Research and Development, LLC, Titusville, NJ

marc.sobel@temple.edu and iturkoz@its.jnj.com

Abstract

Information before unblinding regarding the success of confirmatory clinical trials is highly uncertain. Estimates of expected future power which purport to use this information for purposes of sample size adjustment after given interim points need to reflect this uncertainty. Estimates of future power at later interim points need to track the evolution of the clinical trial. We employ sequential models to describe this evolution. We show that current techniques using point estimates of auxiliary parameters for estimating expected power: (i) fail to describe the range of likely power obtained after the anticipated data are observed, (ii) fail to adjust to different kinds of thresholds, and (iii) fail to adjust to the changing patient population. Our algorithms address each of these shortcomings. We show that the uncertainty arising from clinical trials is characterized by filtering later auxiliary parameters through their earlier counterparts and employing the resulting posterior distribution to estimate power. We devise MCMC-based algorithms to implement sample size adjustments after the first interim point. Bayesian models are designed to implement these adjustments in settings where both hard and soft thresholds for distinguishing the presence of treatment effects are present. Sequential MCMC-based algorithms are devised to implement accurate sample size adjustments for multiple interim points. We apply these suggested algorithms to a depression trial for purposes of illustration.

1. Introduction

- During the design of a confirmatory clinical trial, it is often the case that required information is not fully available and information that is used is often subject to a high degree of uncertainty.
- This information includes, but is not limited to, the expected treatment differences, the assumed population variance, and estimated dropout rates.
- Group sequential and adaptive designs enable the evaluation of uncertainty in the planning phase without compromising the integrity of the trial.
- At interim points during the trial, re-evaluations of preplanned effect sizes and variance estimates may be beneficial. If the original assumptions appear to be incorrect, adjustments can be made to improve the chance that the trial will reach a definitive conclusion. One such adjustment, which has been discussed extensively in the literature, is to modify the sample size (i.e., sample size re-estimation).
- Breaking the blind to perform sample size adjustment in a clinical trial is frequently resource intensive.There are significant credibility issues arising when the sample size is examined using unblinded data. Unblinding may inflate the Type I error rate.
- The 2010 draft guidelines [1] on adaptive designs recommend that:
  - blinded sample size adjustment procedures increase the potential for a successful study while maintaining Type I error control,
  - blinded sample size adjustment procedures greatly reduce the risk of bias, and
  - estimators of variance in support of sample size readjustment are subject to increased variability during the course of the trial.
- ICH guidelines [2] also cover blinded sample size adjustment.
- In view of recommendations (1) and (2), we adopt sample size procedures which provide sample size adjustments in blinded settings. In view of (3), we adopt procedures which: (a) take account of the error resulting from estimating the variance, and (b) adjust to changes in the variance and associated auxiliary parameters over the course of the trial.
- We adopt a Bayesian approach to estimating the variance and associated auxiliary parameters at each stage of the trial; we use particle filter models to adjust for changes in the auxiliary parameters.

2. Overview

There have been three main approaches to sample size determination in general settings in which hierarchical hypotheses are being tested [9] and [10].

(i) Predictive approaches to sample size determination [11],[15], and [16].

(ii) Goal oriented approaches to sample size determination [12], [13], [14], and [17].

(iii) Sample size determination using power estimation. Sample sizes are determined by calculating the "future power" obtainable from adding future observations to the test statistic (used to distinguish whether a significant response is present) [3] and [18]. Historical data have been used in this setting together with the EM algorithm.

We focus on sample size determination using power estimation in blinded settings (item (iii) above). Our methodology can also be applied to goal-oriented sample size determination (item (ii) above, [19], [21], and [22]). We leave this for future work. Posthoc power [20] is the retrospective power of an observed effect based on the sample size and parameter estimates. We compare our results below with those obtained using posthoc power calculations made after the additional subjects have been observed and unblinding has occurred. Gould and Shi [4] calculate power and expected power (using what we refer to below as the approximate strategy) in blinded settings. The approximate strategy fails to:

- provide a range of expected power with regard to what is achievable;
- adjust to the presence of soft treatment effect thresholds; these occur when there is disagreement over which threshold to use; and
- does not adjust to changing patient populations (i.e., the heterogeneity of early versus later enrolled patients).

Our proposed methodology employs a Bayesian strategy to address all of these shortcomings.

- We provide a markov chain monte carlo (MCMC) approach to calculating expected power in both hard and soft threshold settings.
- Particle filters are utilized to formulate models which properly adjust to changing patient populations.
- There are a wide variety of Bayesian strategies proposed in the literature for sample size determination [13], [24], and [27]. Related to these are a number of model selection approaches which employ simulation-based approaches [12], [25], [28], and [34].
- Most Bayesian and model selection strategies involve providing sample size adjustment at a single interim point. The problem of providing a sample size adjustment after a number of interim points have been observed has received much less attention. Below, we offer a sequential framework for addressing this problem.
- We address the issue of how patient population changes between interim points influence sample sizes recommendations by using particle filter methodology [29] and [30].
- Our methodology combines nonsequential Bayesian and model selection strategies for sample size estimation with their sequential counterparts.

3. Previous Work

Gould and Shih ([3] and [4]) discussed modifying the design of ongoing trials without unblinding by providing an adjusted version of the one-sample variance estimator. They proposed a procedure to estimate the within-group variance for sample size re-estimation without unblinding the clinical trial data at interim stages using the EM algorithm. This procedure made use of Maximum Marginal Likelihood Estimates (MMLEs) of within-group variability.
Friede and Kieser [5] and [6] questioned the reliability of the within group variance estimates of the Gould and Shih approach [4] and later provided a number of alternatives for blinded sample size evaluations. Xing [7] used the enrollment order of subjects and the randomization block size to estimate the within group variance.

4. Setup

We propose using information from blinded data. The purpose of this research is to provide a framework for sample size determination under these conditions. We assume two subject groups; our methodology readily extends to more than two groups.

(i) Assume n identical, mutually independent subjects are randomly assigned to the control or experimental treatment groups with known probabilities $1-p$ and p, respectively.

(ii) The parameter δ corresponds to the treatment effect in the clinical trial; θ includes all the auxiliary parameters, such as the pooled standard deviation. The parameters θ and δ are both assumed to be unknown.

(iii) observed subject responses

(a) Observed subject responses X_i in the experimental treatment arm are assumed to be distributed according to $f_1(x|\theta,\delta)$, with known density f_1.

(b) Observed subject responses in the control arm are assumed to be distributed according to $f_0(x|\delta,\theta)$, with known density f_0.

(iv) We use the notation $Z_i = 1$ to indicate that subject i is assigned the treatment; $Z_i = 0$ denotes the control group assignment. The probability that $Z_i = 1$ is assumed to be the known value p (i=1,..,n).

(v) We use the notation $\mathbf{X} = (X_1, ..., X_n)$ for the interim sample having size n. We anticipate that the additional, as yet unobserved, m observations $\mathbf{X}^{(new)} = (X_1^{(new)}, ..., X_m^{(new)})$ can also be selected from the same families of distributions.

(vi) We would like to distinguish between a null and alternative model. We will be concerned with two different settings in which sample size adjustments can be implemented:

(a) In the hypothesis testing setting, tests are devised in which the assumed threshold, distinguishing whether a treatment effect is present, is fixed over the entire length of the trial;

(b) In the model selection setting, more conservative tests with noisy thresholds are devised.

Tests used under a model selection setting are distinguished from those devised in hypothesis testing settings by the assumption of a prior distribution with additional noise for the treatment effect; in this Section we use the notation, λ for the additional (auxiliary) parameters introduced in this case.

(i) Hypothesis Testing Setting:
Deciding which of two disjoint sets (referred to below as g_0 and g_a) the parameter δ belongs to. Group sequential and adaptive designs enable the evaluation of uncertainty in the planning phase without compromising the integrity of the trial. These techniques, including sample size re-estimation have been discussed extensively in the literature. Abusing notation slightly we assume, in this case, that θ has prior $h(\theta)$.

(ii) The Model Selection Setting:
In this setting, the aforementioned models correspond to two distinct families of prior distributions. One model assumes that the parameter of interest δ is distributed according to the family of distributions, $g_0(\bullet|\lambda)$ with unknown (auxiliary) parameter λ. The other model assumes that δ is distributed according to the family of distributions $g_a(\bullet|\lambda)$. We test the hypotheses:

$$H_0 : \delta \sim g_0(\bullet|\lambda)$$
$$H_a : \delta \sim g_a(\bullet|\lambda)$$

where λ is assumed to be independent of the parameter δ but possibly not independent of θ. In view of this, we assume that the parameters θ and λ have joint prior $h(\theta, \lambda)$.

Assuming responses from n subjects are observed, our primary objective is to calculate the additional sample size m required to differentiate between the models g_0 and g_a. Classical statistics interprets this requirement in terms of choosing a number m of additional observations needed to insure that the resulting power of the test distinguishing between the two models is above a particular threshold. We adopt this viewpoint.

5. Theory and Methods

- The proposed strategy assumes that the parameters λ and θ are estimated a posteriori using the data $\mathbf{X} = (X_1, ..., X_n)$. We use simulated data $\mathbf{X}^* = (X_1^*, ..., X_n^*)$ and yet to be observed random variables, $\mathbf{X}^{(new)}$ (of size m) to estimate the power.

- For this purpose we employ conditional likelihood ratio tests ([32] and [33]) and a simulation based approach ([12], [14], [28], and [34]). Our approach makes substantial use of MCMC methodologies [35].
- In the model selection setting, we employ null and alternative posterior distributions ([36] and [37]) which are calculated using marginalized likelihoods. This is frequently equivalent, in the model selection setting, to calculating the posterior probabilities of the null and alternative hypotheses when the simulated and additional data are generated from their (respective) null and alternative predictive posterior distributions [34] and [36]. In the hypothesis testing setting, we construct (generalized) conditional likelihood ratio statistics (CLRT) by choosing parameters least favorable to the null and alternative hypotheses [32] and [33].
- For reasons of convenience, we employ the test value, $T_{n+m}(\mathbf{X}^*, \mathbf{X}^{(\mathbf{new})}, \lambda, \theta, \mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})})$, corresponding to the additive inverse of the conditional likelihood ratio statistic. In hypothesis testing settings, we drop λ from the description of T_{n+m}. The binary treatment assignment variables $\mathbf{Z}^*$ and $\mathbf{Z}^{(new)}$, associated respectively with the simulated and as yet unobserved values are assumed to have their prior distribution: $P(Z = 1) = p$. For purposes of simplification, we assume that p is 0.5 and that the experimental design has two treatment groups. Our results generalize easily to arbitrary p with more than two treatment groups. In both the critical value and power calculations described below, we make extensive use of the law of large numbers and the central limit theorem [31]. We adopt the notation $(\bullet)$ for the density of the variable $\bullet$. The notation, P_H denotes the posterior probability operator over the simulated and anticipated observations $\mathbf{X}^*, \mathbf{X}^{(\mathbf{new})}$ under the null and alternative hypotheses $H = H_0, H_a$, respectively.
- The critical value of the test is a parameter $Crit = Crit(\lambda, \theta)$ whose posterior distribution is calculated with $Crit = Crit(\lambda, \theta)$ and $T_{n+m} = T_{n+m}(\mathbf{X}^*, \mathbf{X}^{(\mathbf{new})}, \lambda, \theta, \mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})}))$ from:

$$\begin{aligned} \alpha &= P_{H_0}\left(T_{n+m} < Crit \middle| \lambda, \theta\right) \\ (\mathbf{X}^*, \mathbf{X}^{(new)}) &\propto m_{H_0}(\bullet|\lambda, \theta, \mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})})(\mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})}) \\ \left(\lambda, \theta \middle| \mathbf{X}\right) &\propto m_{H_0}(\mathbf{X}|\lambda, \theta, Z)h(\theta, \lambda)(Z) \end{aligned} \quad (1)$$

The notation $m_H(\bullet|\lambda, \ldots)$ used in equations (1) and (2), refers to the distribution of the observations (both simulated and anticipated) marginalized over the hypothesis $H = H_0, H_a$; the notation $m_H(X|\lambda, \theta, \mathbf{Z})$ refers to the observations marginalized over the hypothesis $H = H_0, H_a$.

- The power of the test, also a parameter, is then calculated using the previously calculated critical parameter and the anticipated observations $\mathbf{X}^{(\mathbf{new})}$ together with the above specifications via:

$$\begin{aligned} \text{Power}(n+m|\lambda, \theta) &= P_{H_a}(T_{n+m} < Crit) \\ (\mathbf{X}^{(new)}, \mathbf{X}^*) &\propto m_{H_a}(\bullet|\lambda, \theta, \mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})})(\mathbf{Z}^*, \mathbf{Z}^{(\mathbf{new})}) \\ \left(\lambda, \theta \middle| \mathbf{X}\right) &\propto m_{H_a}(\mathbf{X}|\lambda, \theta, \mathbf{Z})h(\lambda, \theta)(\mathbf{Z}) \end{aligned} \quad (2)$$

- Having simulated the power aposteriori, we calculate High Posterior Density (HPD) intervals for the power and present them in lieu of fixed power estimates. The critical value given in equation (1) and the power given in equation (2) are analogous to the (Bayesian) sample size determination (SSD) separation quantities given in Wang and Gelfand (equations 10a and 10b of [34]).

6. Relevant Theoretical Results

The primary theorems relevant to using our formulation are the following: We use the notation $\widehat{\delta}_{n+m}[H] = \widehat{\delta}_{n+m}(\lambda, \theta)[H]$ for the value of δ resulting from maximizing the likelihood combining the simulated and additional observations under the given hypothesis $H = H_0, H_a$. To simplify notation, we do not drop the auxiliary parameter λ from discussion of the hypothesis testing setting, below. In accordance with this simplification, $I(X)$ denotes a given posterior HPD interval for the parameters λ and θ.

Theorem 1 *The critical value parameter $crit = crit(\lambda, \theta)$ can be chosen to satisfy the significance level condition,*

$$P_{H_0}\left(T_{n+m}(\mathbf{X}^*, \mathbf{X}^{(\mathbf{new})}, \lambda, \theta) < crit\right) = \alpha \quad (3)$$

if, for the conditional HPD interval $\lambda, \theta \in I(X)$ having given size, eventually as $m \to \infty$:

$$\widehat{\delta}_{n+m}(\lambda, \theta)[H_0] \in g_0, \forall \lambda, \theta \in I(X) \quad (4)$$

or, in the model selection setting, $\forall \epsilon > 0$

$$\lim_{m\to\infty} P_{\delta\sim g_0}\left(|\widehat{\delta}_{n+m}(\lambda, \theta)[H_0] - \delta| < \epsilon | X\right) = 1, \quad \forall \lambda, \theta \in I(X) \quad (5)$$

Theorem 2 *Using the same notation as was introduced in Theorem 1, and assuming that the critical value parameter $crit = crit(\lambda, \theta)$ has been chosen, the condition,*

$$P_{H_a}\left(T_{n+m}(\mathbf{X}^*, \mathbf{X}^{(\mathbf{new})}, \lambda, \theta) < crit\right) \geq 1 - \beta \quad (6)$$

is satisfied for large enough m if, for the conditional HPD interval $\lambda, \theta \in I(X)$, having given size, eventually as m tends to infinity,

$$\widehat{\delta}_{n+m}(\lambda, \theta)[H_a] \in g_a, \forall \lambda, \theta \in I(X) \quad (7)$$

or, in the model selection setting, $\forall \epsilon > 0$,

$$\lim_{m\to\infty} P_{\delta\sim g_a, \lambda, \theta|X}\left(|\widehat{\delta}_{n+m}(\lambda, \theta)[H_a] - \delta| < \epsilon | X\right) = 1 \quad (8)$$

Theorems 1 and 2 hold in the hypothesis testing setting if, for example, the assumed prior for θ is proper and supported on the whole real line. The theorems hold in the model selection setting if, in addition to the aforementioned assumption, the assumed prior for the λ is proper and has conditional support (for all θ) on the whole real line.

7. Hypothesis Testing Setting: Depression Trial Example

- Assume n identical, mutually independent subject responses, $\mathbf{X} = (X_1, ..., X_n)$ are observed at the first interim stage. Each subject is randomly assigned to the control or experimental treatment groups with known probabilities $1 - p$ and p, respectively. This setting can easily be adapted to more than two treatment groups. We assume that lower response scores indicate improvement.
- The average effect of the treatment is denoted by the parameter δ; the mean control response is denoted by the auxiliary parameter μ. The pooled standard deviation is denoted by the auxiliary parameter τ. Using the notation of Section 5:

(i) the auxiliary parameter θ corresponds to (μ, τ),
- subject responses in the treatment arm are distributed according to $f_1(\bullet|\delta, \theta) = \mathcal{N}(\mu - \delta, \tau)$ and
- subject responses in the control arm are distributed according to $f_0(\bullet|\delta, \theta) = \mathcal{N}(\mu, \tau)$.

(ii) We test the hypothesis,

$$H_0 : \delta = 0; \quad (9)$$
$$H_a : \delta > \delta_1 \quad (10)$$

The latent variable Z_i is 1 if the i'th subject is in the treatment arm and 0 otherwise. For simplicity we assume two treatment groups with a 1:1 allocation ratio; thus p=0.5.

(iii) We employ the quantity,

$$T(X, Z, \tau, \mu) = \frac{\sum_i (X_i - \mu) Z_i}{\tau^2} \quad (11)$$

which is the additive inverse of the conditional likelihood ratio, up to a constant of proportionality.

(iv) We calculate the posterior distributions under the null and separately under the alternative.

(v) Parameters calculated under the null posterior are denoted by: μ_0 and τ_0; those under the alternative posterior are denoted by: μ_a and τ_a. We assume τ^2 has an (an approximately indifferent) inverse gamma prior with shape hyperparameter 1 and (a small) scale hyperparameter ϵ_1.

The proposed approach:
For a given significance level ($\alpha = .05$), the critical parameter c is necessary to compute the power associated with m additional observations. We characterize the posterior distribution of c using B MCMC simulations indexed by $b = 1, ..., B$ of the marginal null posterior distribution. We characterize the posterior distribution of the power using B analogous simulations of the marginal alternative posterior distributions. Below, Φ denotes the standard normal cdf.

$$c^{(b)} = \frac{\Phi^{-1}(\alpha)\sqrt{n+m}}{\tau_0^{(b)}\sqrt{2}} \quad (12)$$

$$\text{power}^{(b)}[n+m] = \Phi\left\{\left[\frac{\tau_a^{(b)}}{\tau_0^{(b)}}\Phi^{-1}(\alpha) + \frac{\delta_1}{\tau_a^{(b)}}\left(\sqrt{\frac{n+m}{2}}\right)\right]\right\} \quad (13)$$

Power is properly estimated by an HPD interval taking the form:

$$\underline{\text{power}}[n+m] < \text{power} < \overline{\text{power}}[n+m]$$

where $\underline{\text{power}}[n+m]$ denotes a lower posterior quantile and $\overline{\text{power}}[n+m]$ an upper posterior quantile measurement. We employ HPD power estimates as described in equation (7). In order to apply the proposed methods, we considered a placebo-controlled study of depression. The details of this trial are given in Mahmoud et al. [38]. Adult outpatients with major depressive disorders who had an incomplete response to antidepressant treatment were randomly assigned (1:1) to active drug or placebo regimens for 6 weeks duration in a double-blind multicenter trial. The primary efficacy endpoint was the mean difference between treatments at endpoint using a 17-item Hamilton Rating Scale for Depression (HRSD-17). A sample size of 116 patients in each group was anticipated to have 90% power to detect a difference in mean HRSD-17 total score change from baseline of 3.0 units assuming that the common standard deviation was 7 using a two-group t-test with a 0.05 two-sided significance level. Adjusting for drop outs, approximately 270 subjects were assumed to be randomized.
Enrollment visit dates were used to order subject entry into the trial. Sample size assumptions were evaluated for demonstration purposes after the 100$^{\text{th}}$, 150$^{\text{th}}$, 200$^{\text{th}}$, and 250$^{\text{th}}$ subject completed the trial. In the left panel of Figure 1, 90% HPD intervals were calculated for expected power after the 100$^{\text{th}}$ subject level data (i.e.,observation) had been examined. The posterior null and alternative distributions for the hypothesis testing (respectively, model selection setting) of the pooled standard deviation τ are given in the left (respectively, right) panels of Figure 2; their mean corresponds roughly to the pooled standard deviation estimates computed in the original study. In the next Section, we examine multiple stage sample size determination in the context of this example.

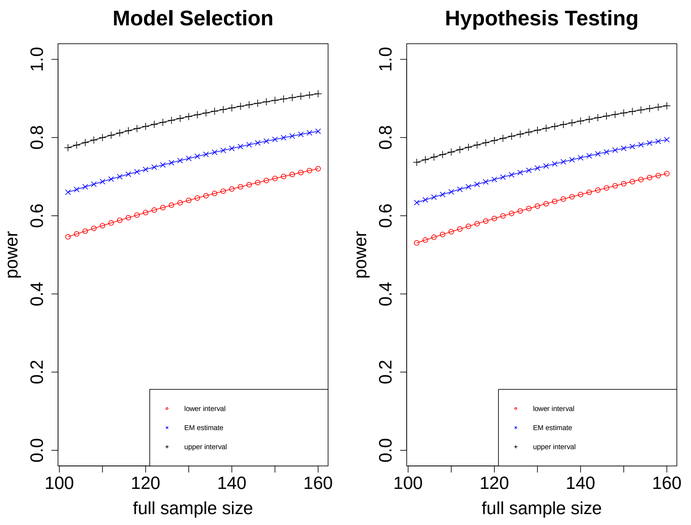

Figure 1: *90 Percent HPD intervals for expected power using the first 100 subjects when 50 additional observations are anticipated for the Depression Trial in the hypothesis testing and model selection settings, respectively. Posthoc power was calculated to be 66% after the 150th observation. This is well within the HPD interval given in the left panel.*

Data Set	N, LS-Mean (SE) Drug	N, LS-Mean (SE) Placebo	LS-Mean Diff (SE), [95% CI]	Post-Hoc Power at 0.05	Post-Hoc Power at 0.025
First 100 subjects	54, 16.1 (1.0)	46, 19.3 (1.0)	-3.2 (1.4), [-6.0 ;-0.4]	0.610	0.496
101-150 subjects	29, 15.9 (1.3)	21, 17.6 (1.5)	-1.8 (2.0), [-5.8 ; 2.3]	0.138	0.084
First 150 subjects	83, 16.0 (0.8)	67, 18.8 (0.8)	-2.7 (1.1), [-5.0 ;-0.5]	0.663	0.553
151_200 subjects	23, 13.1 (1.2)	27, 16.2 (1.1)	-3.1 (1.6), [-6.4 ; 0.2]	0.462	0.349
First 200 subjects	106, 15.4 (0.7)	94, 18.0 (0.7)	-2.6 (1.0), [-4.5 ;-0.8]	0.784	0.692
201_258 subjects	26, 16.5 (1.5)	32, 16.0 (1.4)	0.5 (2.0), [-3.5 ; 4.5]	0.056	0.029
Complete Data Set	132, 15.6 (0.6)	126, 17.5 (0.6)	-1.9 (0.9), [-3.6 ;-0.2]	0.591	0.478

Table 1: *Posthoc power, calculated in an unblinded setting, for the Depression data*

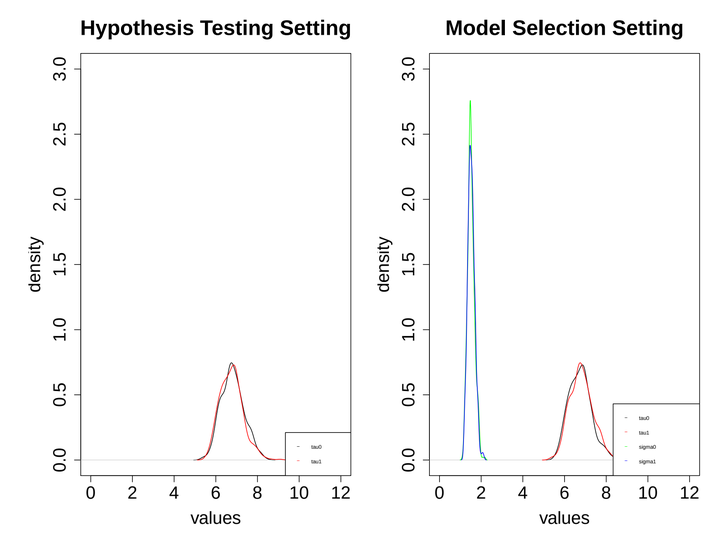

Figure 2: *Null and Alternative posterior densities of the τ and σ parameters in the model selection and hypothesis testing setting*

8. Clinical Trial Sample Size Adjustments in a Model Selection Setting

- The model selection setting is similar to that given in Section 7, but we now incorporate a variety of judgments about the threshold δ_1, distinguishing whether a treatment effect is present.
- By adding noise to both the null and alternative hypotheses, we effectively incorporate all of these judgments; we call this the, "model selection setting." Tests in model selection settings are more conservative and hence give rise to smaller expected power than their hypothesis testing counterparts.
- The average effect of the treatment is denoted by the parameter δ; the mean control response is denoted by μ. The pooled standard deviation is denoted by τ. Subjects in the treatment arm are assumed to be distributed according to $\mathcal{N}(\mu - \delta, \tau)$; subjects in the control arm are

distributed according to $\mathcal{N}(\mu, \tau)$. Our objective in this Section is to test the noisy (vague) null and alternative hypotheses given below.

- The null and alternative hypotheses, for fixed, known δ_1, are:

$$H_0 : \delta \sim \mathcal{N}(0, \sigma^2) \quad (14)$$
$$H_a : \delta \sim \mathcal{N}(\delta_1, \sigma^2) \quad (15)$$

- The null hypothesis effectively adds the (Gaussian) noise factor $\mathcal{N}(0, \sigma^2)$ to the null hypothesis assumed in equation (9) and the alternative hypothesis in equation (10).
- The addition of noise converts the assumed hard threshold δ_1 into a soft threshold.
- In the notation of Section 5, the auxiliary parameter λ corresponds to the parameter σ, defined above.
- The marginal likelihoods (ML) under the null and alternative hypotheses are:

$$\text{ML under null} \sim \frac{\exp\left\{-(1/2)\sum_i \left(\frac{(X_i-\mu)^2}{Z_i\sigma^2+\tau^2}\right)\right\}}{\prod_i \sqrt{(Z_i\sigma^2+\tau^2)}}$$

$$\text{ML under alternative} \sim \frac{\exp\left\{-(1/2)\sum_i \left(\frac{(X_i-\mu+\delta_1 Z_i)^2}{Z_i\sigma^2+\tau^2}\right)\right\}}{\prod_i \sqrt{(Z_i\sigma^2+\tau^2)}}$$

- Let $\psi = \sigma^2 + \tau^2$. We assume a nearly indifferent prior for ψ, the usual Bernoulli prior $(p,\ 1-p)$ for the Z's, and the prior described in Section 7 for τ^2. We assume an inverse gamma prior for ψ having shape parameter 1 and scale ϵ_1. σ^2 inherits a prior from that given for τ^2 and ψ.
- We can compute the critical value parameter by first marginalizing over the null and separately over the alternative hypotheses (see e.g., [36]).
- The additive inverse of the conditional likelihood ratio statistic is:

$$T_n(\mathbf{X}, \mu, \sigma, \tau, \mathbf{Z}) \propto \sum_i \left(\frac{Z_i(X_i-\mu)^2 - Z_i(X_i-\mu+\delta_1)^2}{Z_i\sigma^2+\tau^2}\right)$$
$$\propto \frac{\sum_{i=1}^n Z_i(X_i-\mu)}{\psi}$$

- The proposed approach: We adopt the same conventions as were adopted in section 7 (above). The notation ψ, is as defined above. The quantities, ψ_0 and ψ_a, denote the parameter ψ under the null and alternative posterior distributions, respectively. The critical value and power parameters are computed at significance level α as:

$$c^{(b)} = \frac{\Phi^{-1}(\alpha)\sqrt{n+m}}{\sqrt{\psi_0^{(b)}}\sqrt{2}} \quad (16)$$

$$\text{power}^{(b)}[n+m] = \Phi\left\{\left[\frac{\sqrt{\psi_a^{(b)}}}{\sqrt{\psi_0^{(b)}}}\Phi^{-1}(\alpha) + \frac{\delta_1}{\sqrt{\psi_a^{(b)}}}\left(\sqrt{\frac{n+m}{2}}\right)\right]\right\} \quad (17)$$

- As an example, we describe our results for the depression study. The parameters τ and σ take on a variety of values in this case under both the null and alternative posterior distributions, as a consequence of the noisy nature of the test (see Figure 2).
- Note the lower expected power in this case. This is a consequence of the fact that the hypotheses are noisier and hence provide less evidence of future power. (see Figure 1, right panel).

9. Advanced Stage Sample Size Determination

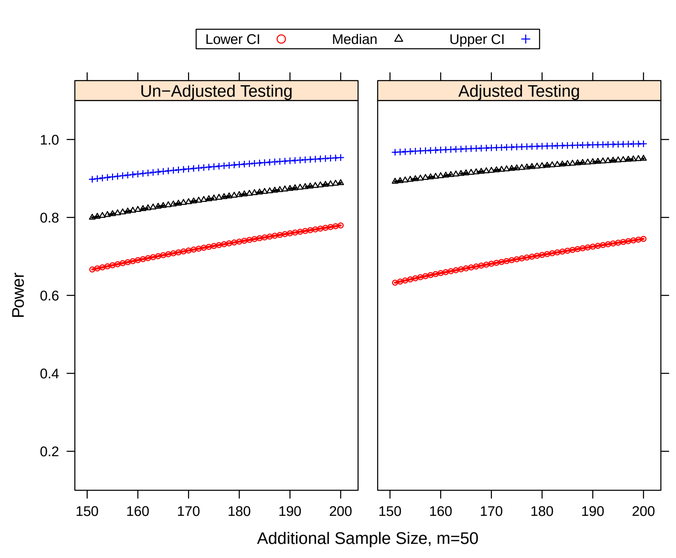

Figure 3: *90 percent HPD intervals for the Advanced Stage Power Estimation after Samples of sizes 100+50 have been observed. The figure on the left estimates power in the Depression Trial assuming a change in reliability; the figure on the right estimates power assuming no change in reliability. Table 1 gives a (posthoc) power of 78% after 200 observations. This is roughly comparable to the lower HPD quantile but not the median HPD quantile.*

- Nearly all ongoing clinical trials are monitored continuously and blinded data sets become available at prespecified periodic intervals. This condition provides ample opportunity to examine blinded data at various interim points.
- Early enrolled patients frequently demonstrate different behavior than those patients entering the study later.
- In the depression trial, introduced above, early enrolled patients demonstrated more reliable behavior than those entering the study later. In this case, predictions are improved by giving more weight to earlier patients.
- In this Section, we propose an algorithm for calculating sample size adjustments at a later interim point which takes account of the aforementioned reliability concerns. This enables us to accurately update the auxiliary parameters using all of the data observed before the adjustment is recommended. We refer to this below as advanced stage sample size adjustment.
- We demonstrate our results for two interim points; generalizations to more than two interim points are clear.

Using notation analogous to that introduced in Section 5:

(i) Assume at interim stage j (j=1,2), n_j identical, mutually independent subjects are randomly assigned to treatment groups with known probability, p.

(ii)
- Subjects in the experimental treatment arm with observed values $X_{i,j}$ (i=1,..,n; j=1,2) are modeled as coming from the normal distribution $\mathcal{N}(\mu - \delta Z_{i,j}, \tau_j)$.
- Subjects in the control arm with observed values $X_{i,j}$ are modeled as coming from the normal distribution, $\mathcal{N}(\mu, \tau_j)$.

(iii) We use the notation $Z_{i,j} = 1$ to indicate that subject i corresponding to interim stage j is assigned the treatment; we use the notation $Z_{i,j} = 0$ to denote the control group assignment. The probability of $Z_{i,j} = 1$ is assumed to be p. We use the notation $\mathbf{X_j} = (X_{1,j}, ..., X_{n_j,j})$ for the interim sample having size n_j (j=1,2).

(iv) We anticipate that the additional, as yet unobserved, m observations $\mathbf{X}^{(new)} = \mathbf{X_{n+1:m}} = (X_{n+1}, ..., X_m)$ are generated from the normal distribution $\mathcal{N}(\mu - \delta Z_{i,j}, \tau_2)$. We test the null and alternative hypotheses given by

$$H_0 : \delta = 0 \quad (18)$$
$$H_a : \delta > \delta_1; \quad (19)$$

(v) We use the notation $f_{h,j}$ to denote the likelihood under hypothesis $h = 0, a$ at interim point j and $\tau_{h,j}$ for the scale parameter under hypothesis h at interim point j. We omit mention of μ in this notation.

(vi) κ_{h1} and κ_{h2} characterize the shape and scale respectively of gamma random variables tending to take values larger than 1 having a ratio larger than 1.

(vii) The gamma variable c_h with shape κ_{h1} and scale κ_{h2}, used below, reflects the presumed reduction in certainty in going from the first to the second interim data set; we use the notation $Gamma(\kappa_{h1}, \kappa_{h2})$ for the resulting gamma distribution. We assume the same approximate indifference prior for $\tau_{h,1}$ as was assumed for τ_h above.

(viii) We employ the model:

$$\begin{aligned} \mathbf{X_1} &\sim f_{h,1}(\bullet|\tau_{h,1}) \\ \tau_{h,2} &\sim c_h\tau_{h,1} \qquad c_h \sim Gamma(\kappa_{h1}, \kappa_{h2}) \\ \mathbf{X_2} &\sim f_{h,2}(\bullet|\tau_{h,2}) \end{aligned} \quad (20)$$

(ix) Note that, by assumption, the prior distribution for $\tau_{h,2}$ provides a large prior probability that $\tau_{h,2}$ is larger than $\tau_{h,1}$; the size of this probability depends on the hyperparameters κ_{h1} and κ_{h2} for $h = 0, a$. We adopt the notation $(\bullet|\mathbf{X}_1, \mathbf{X}_2)_{H_j}$ to denote the H_j posterior distribution of τ given all of the observed data.

(x) Posterior inference in this case makes use of standard particle filter algorithms [29] and [30].

(xi) We calculate critical values and power using the posterior distributions:

$$\tau_{0,2}^{(b)} \sim (\bullet|\mathbf{X}_1, \mathbf{X}_2)_{H_0}; \quad b = 1, ..., B$$

$$\tau_{a,2}^{(b)} \sim (\bullet|\mathbf{X}_1, \mathbf{X}_2)_{H_a}; \quad b = 1, ..., B$$

(xii) The critical values and power can then be calculated using:

$$c^{(b)} = \frac{\Phi^{-1}(\alpha)\sqrt{n+m}}{\tau_{0,2}^{(b)}\sqrt{2}} \quad (21)$$

$$\text{power}^{(b)}[n+m] = \Phi\left\{\left[\frac{\tau_{a,2}^{(b)}}{\tau_{0,2}^{(b)}}\Phi^{-1}(\alpha) + \frac{\delta_1}{\tau_{a,2}^{(b)}}\left(\sqrt{\frac{n+m}{2}}\right)\right]\right\} \quad (22)$$

(for $n = n_1 + n_2$).

(xiii) We calculate estimated power using the data from the depression trial, described previously. The first interim point comes after 100 data points are observed; the second comes after an additional 50 data points have been observed. We assumed $\kappa_{h1} = 3$ and $\kappa_{h2} = 2$. Our results were compared with those for which no change in reliability was assumed (i.e., the original framework). The upper and median quantiles of the adjusted future power estimates given in Figure 3 are comparable to their unadjusted counterparts. The lower quantiles of the adjusted future power estimate is substantially smaller than its unadjusted counterpart; this is a consequence of the fact that by adjusting for the greater reliability of earlier patients we give less weight to the accumulated evidence against the null at interim point 2.

We note that Table 1 gives a (posthoc) power of .78 after 200 observations. This is easily within the scope of the adjusted HPD interval but roughly outside the scope of the unadjusted HPD interval (see Figure 3).

10. Conclusion

(i) We have argued in favor of:
- (a) providing sample size adjustments before unblinding,
- (b) providing adjustments in both soft and hard threshold settings, and
- (c) providing more accurate and more flexible auxiliary parameter (e.g., variance) estimators in support of changes in patient population.

(ii) In further support we note the many guidelines recommending these adjustment changes (see Section 1).

(iii) The information available before unblinding, although useful, is highly uncertain. Estimates of expected (future) power obtained in this setting need to reflect this uncertainty. We have shown that current techniques using point estimates of auxiliary parameters for estimating expected power fail to:
- (a) accurately describe the range of likely power obtained after the anticipated data are observed,
- (b) fail to anticipate the need for sample size adjustments in the presence of both hard and soft threshold settings, and
- (c) fail to adjust to changes in the patient population.

(iv) The procedures devised above addressed all of these shortcomings.

(v) Breaking the blind to perform sample size adjustment in a clinical trial is resource intensive; blinded sample size re-estimation is generally well accepted by regulators. Nearly all ongoing clinical trials are monitored continuously and data sets become available at periodic intervals. This monitoring provides ample opportunity to examine blinded data at various interim points. The data set consisting of the collection of all interim data sets is the *combined* data set. Patients enrolled in earlier interim data sets may demonstrate more reliable behavior than patients entering the study at a later point.

(vi) The proposed multistage algorithm provides flexibility in assigning weights to auxiliary parameters associated with different interim data points, according to the subjective assessment of the researcher.

(vii) For the depression example, predictions are frequently more accurate when the pooled standard deviation for early enrolling patients is assumed to be smaller apriori than the pooled standard deviation for their later enrolling counterparts. We have argued that this difference in accuracy should be modeled by filtering auxiliary parameters arising from later interim points through those arising from earlier interim points.

(viii) Particle filter models were shown to provide an appropriate mechanism for modelling these prior relationships.

(ix) In the depression trial example, the differences in response between the last set of subjects and the first 200 subjects were apparent. This response heterogeneity had a significant effect on the posthoc power, underscoring the need for estimates of future power which accurately model it. The uncertainty in the information available before unblinding is accurately characterized by statistical models which make use of the posterior distribution, conditional on the observed response data, of the auxiliary parameters. More generally, response heterogeneity over the course of a clinical trial, is a common problem; it is hoped that the suggested methodology can be useful.

References

[1] (2010) FDA Guidelines, web address: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf .

[2] (2010) ICH Guidelines, web address: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073137.pdf .

[3] Gould AL, Shih WJ. (1998) Modifying the design of ongoing trials without unblinding. *Statistics in Medicine*, 17, pp. 89-100.

[4] Gould AL, Shih WJ. (1992) Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance, *Communications in Statistics Theory and Methods*, 21, pp. 2833-2853.

[5] Friede T, Kieser M. (2002) On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine*, 21, pp. 165-176.

[6] Friede T, Kieser M. (2003) Blinded sample size assessment in non-inferiority and equivalence trials. *Statistics in Medicine*, 22, pp. 995-1007.

[7] Xing B, Ganju J. (2005) A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, 24: 1807-1814

[8] Kieser M, Friede T., (2003) Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, 22, pp. 3571-3581.

[9] Peseshk, H. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods for Medical Research* 12, pp. 489-504.

[10] Adcock, C.J., (1997) Sample Size Determination: A Review. *The Statistician*, 46.

[11] Zhong,W., et. al. (2013) A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. Contemp Clin Trials. Nov;36(2):587-596

[12] Santis, F.D. and Spezzaferri, F. (1997) Alternative Bayes factors for model selection. *The Canadian Journal of Statistics*, 25, pp 503-515.

[13] Sahu,S.K., and Smith, T.M.F. (2006) A Bayesian method for sample size determination with practical applications. *J.R. Statist. Soc. A* 169, pp. 235-253.

[14] Santis, F.D. (2007) Using historical data for Bayesian sample size determination. *J.R. Statist. Soc. A.*, 170, pp. 95-113.

[15] Geisser, S., and Eddy, W., (1979) A predictive approach to model selection, *J. Amer. Statist. Assoc.*, 74, pp 153-160.

[16] Lee, S.J., and Zelen, M. (2000) Clinical Trials and Sample Size Considerations: Another Perspective *Statistical Science*, 15, pp. 95-110.

[17] Inoue, L.Y.T., et. al. (2005) Relationship between Bayesian and Frequentist Sample Size Determination. *The American Statistician*, 59, pp.79-87.

[18] Santis, F.D., and Spezzaferri, F., (1997) Alternative Bayes Factors for Model Selection. *The Canadian Journal of Statistics*, 25. pp. 503-515.

[19] Schuster, J.J. (1993) *Practical Handbook of Sample Size Guidelines for Clinical Trials*, CRC Press, Boca Raton, Fl.

[20] Lenth, R. (2013) Posthoc Power: Tables and Commentary. *Technical Report 368, Department of Statistics and Accounting, University of Iowa*, Available on the web.

[21] Self, S.G. and Mauritsen, R.H., (1988) Power/sample size calculations for generalized linear models. *Biometrics*, 44, pp 79-86.

[22] Self, S.G. and Mauritsen, R.H., (1992) Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48, pp 31-39.

[23] Aitken, M. (1991) Posterior Bayes Factors. *J.R. Statist. Soc. B*, 53, pp.111-142.

[24] Joseph,L. and Belisle, P., (1997) Bayesian sample size determination for normal means and differences between normal means. *The Statistician*, 46, pp. 209-226.

[25] O'hagan, (1995) Anthony, Fractional Bayes Factors for Model Comparison, *Journal of the Royal Statistical Society B*, pp. 99-138

[26] Belisle, P. and Joseph, L. (1997) Bayesian sample size determination for normal means and differences between normal means. *The Statistician*, 46, pp. 208-226.

[27] Hartley, A. Adaptive blinded sample size adjustment for comparing two normal means - a mostly Bayesian approach, *Pharmaceutical Statistics*, 2012, 11, pp. 230-240.

[28] Gelfand, A., and Dey, D.K., (1994) Bayesian model choice: Asymptotics and Exact calculation. *Journal of the Royal Statistical Soc. Ser. B*, 56, pp. 501-506.

[29] Doucet, A.,Godsill, S., and Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing*, 10, pp. 197-208.

[30] Carvalho, C., et al. (2010) Particle Learning and Smoothing, *Statistical Science*, 2010, 25, pp. 88-106.

[31] Gnedenko, B.V. (1969) *The Theory of Probability*, Mir Publishers.

[32] Lehmann, E. (1986) *Testing Statistical Hypotheses*, Wiley, Second Edition, New York.

[33] Meng, X.Li (1994) Posterior Predictive p-Values. *The Annals of Statistics*, 22, pp. 1142-1160.

[34] Wang, F., and Gelfand A. (2002) A Simulation-based Approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17, pp. 193-208.

[35] Robert, C., and Cassella, G., (2004) *Monte Carlo Statistical Methods*, Springer, Second Edition.

[36] Weiss, R., (1997) Bayesian sample size calculations for hypothesis testing. *The statistician*, 46.

[37] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit *J. Roy. Statist. Soc. Ser. B*.

[38] Mahmoud, R.A., et al. (2007) Risperidone for Treatment-Refractory Major Depressive Disorder: A Randomized Trial, *Ann Intern Med.*, 147(9), pp 593-602.