BLINDED EVALUATIONS OF EFFECT SIZES IN CLINICAL TRIALS: COMPARISONS BETWEEN BAYESIAN AND EM ANALYSES

A Dissertation Submitted to the Temple University Graduate Board in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

By
Ibrahim Turkoz

May 2013

Examining Committee Members:
Marc Sobel, PhD; Advisory Chair; TU Department of Statistics
Richard M. Heiberger, PhD; TU Department of Statistics
Yuexiao Dong, PhD; TU Department of Statistics
Zhigen Zhao, PhD; TU Department of Statistics
Jose Pinheiro, PhD; External Member; Senior Director of Statistics at Janssen Research & Development LLC, Johnson & Johnson Company

ABSTRACT

Clinical trials are major and costly undertakings for researchers. Planning a clinical trial involves careful selection of the primary and secondary efficacy endpoints. The 2010 draft FDA guidance on adaptive designs acknowledges possible study design modifications, such as selection and/or order of secondary endpoints, in addition to sample size re-estimation. It is essential for the integrity of a double-blind clinical trial that individual treatment allocation of patients remains unknown. Methods have been proposed for re-estimating the sample size of clinical trials, without unblinding treatment arms, for both categorical and continuous outcomes. Procedures that allow a blinded estimation of the treatment effect, using knowledge of trial operational characteristics, have been suggested in the literature.

Clinical trials are designed to evaluate effects of one or more treatments on multiple primary and secondary endpoints. The multiplicity issues when there is more than one endpoint require careful consideration for controlling the Type I error rate. A wide variety of multiplicity approaches are available to ensure that the probability of making a Type I error is controlled within acceptable pre-specified bounds. The widely used fixed sequence gate-keeping procedures require prospective ordering of null hypotheses for secondary endpoints. This prospective ordering is often based on a number of untested assumptions about expected treatment differences, the assumed population variance, and estimated dropout rates.

We wish to update the ordering of the null hypotheses based on estimating standardized treatment effects. We show how to do so while the study is ongoing, without unblinding the treatments, without losing the validity of the testing procedure, and with maintaining the integrity of the trial. Our simulations show that we can reliably order the standardized treatment effect also known as signal-to-noise ratio (δ / σ), even though we are unable to estimate the unstandardized treatment effect (δ).

In order to estimate treatment difference in a blinded setting, we must define a latent variable substituting for the unknown treatment assignment. Approaches that employ the EM algorithm to estimate treatment differences in blinded settings do not provide reliable conclusions about ordering the null hypotheses. We developed Bayesian approaches that enable us to order secondary null hypotheses. These approaches are based on posterior estimation of signal-to-noise ratios (δ / σ). We demonstrate with simulation studies that our Bayesian algorithms perform better than existing EM algorithm counterparts for ordering effect sizes. Introducing informative priors for the latent variables, in settings where the EM algorithm has been used, typically improves the accuracy of parameter estimation in effect size ordering. We illustrate our method with a secondary analysis of a longitudinal study of depression.

ACKNOWLEDGMENTS

I would like to express my appreciation to my committee members, my colleagues, and my family. Without their guidance, persistent help, and support this dissertation would not have been possible.

A special thanks and deepest appreciation to my advisor, Dr. Marc Sobel, for the patient guidance and mentorship he provided to me through completion of this research. I am truly fortunate to have had the opportunity to work with Dr. Sobel. I would like to thank my committee members Dr. Richard M. Heiberger, Dr. Yuexiao Dong, Dr. Zhigen Zhao, and Dr. Jose Pinheiro for their insight, excellent guidance, and thought provoking suggestions. I would also like to thank Dr. Jagbir Singh for his encouraging words and wonderful support.

I would like to acknowledge the many years of support from my colleagues and management at J&J. A special thanks to Dr. Steve Ascher and Dr. William Olson for sharing their enthusiasm for and comments on my work.

Finally, I would like to thank my family and friends for their unconditional love and support. I’d be remiss if I didn’t acknowledge the countless sacrifices made by my wife, Heather, in shouldering far more than her fair share of the parenting.

DEDICATION

To future scientists Hannah and Arman for asking very difficult questions every passing year and my parents Erdogan and Senem for their unwavering support.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS v

DEDICATION vi

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1 INTRODUCTION 1

1.1 Background 1

1.2 Literature Review 7

1.3 Objective 13

CHAPTER 2 METHODS 16

2.1 EM Algorithm 18

2.2 Bayesian Approach 20

2.3 Bayesian Approach with Covariate 22

2.4 Bayesian Approach with Covariate and a Power Prior 25

2.5 Model Selection 28

2.6 MCMC: Gibbs Sampling 29

CHAPTER 3 SIMULATIONS 31

3.1 Simulation Set-up 32

3.2 Simulation Results 34

CHAPTER 4 APPLICATION 40

CHAPTER 5 DISCUSSION AND CONCLUSION 47

CHAPTER 6 FUTURE RESEARCH 50

REFERENCES CITED...54

APPENDICES 58

APPENDIX A SIMULATION DETAILS 58

APPENDIX B SIMULATION RESULTS, SCENARIO 1.............................61

APPENDIX C SIMULATION RESULTS, SCENARIO 2.............................67

APPENDIX D SIMULATION RESULTS, SCENARIO 3.............................72

APPENDIX E SIMULATION RESULTS, SCENARIO 4.............................77

APPENDIX F SIMULATION RESULTS, SCENARIO 5 82

APPENDIX G MDD STUDY RESULTS...87

LIST OF TABLES

Table 1. MDD Study Results at Week 4 and Week 6 for HRSD-17 and PaRTS-D Scales Using ANCOVA Models 42

Table 2. MDD Study Simulation Results – DIC Scores on HRSD-17 and PaRTS-D...... 46

Table 3. Scenario 1, Input Data Set Parameter Characteristics 61

Table 4. Scenario 1, Effect Size Ordering of Input Data Sets 62

Table 5. Scenario 1, Summary of Posterior Means of Parameters Over 20 Data Sets 65

Table 6. Scenario 2, Input Data Set Parameter Characteristics 67

Table 7. Scenario 2, Effect Size Ordering of Input Data Sets 67

Table 8. Scenario 2, Summary of Posterior Means of Parameters Over 20 Data Sets 70

Table 9. Scenario 3, Input Data Set Parameter Characteristics 72

Table 10.Scenario 3, Effect Size Ordering of Input Data Sets 72

Table 11. Scenario 3, Summary of Posterior Means of Parameters Over 20 Data Sets ... 75

Table 12. Scenario 4, Input Data Set Parameter Characteristics 77

Table 13. Scenario 4, Effect Size Ordering of Input Data Sets 77

Table 14. Scenario 4, Summary of Posterior Means of Parameters Over 20 Data Sets ... 80

Table 15. Scenario 5, Input Data Set Parameter Characteristics 82

Table 16.Scenario 5, Effect Size Ordering of Input Data Sets 82

Table 17. Scenario 5, Summary of Posterior Means of Parameters Over 20 Data Sets ... 85

LIST OF FIGURES

Figure 1. Posterior Probability of Effect Size Ordering by Scenario 36

Figure 2. Posterior Distribution of Effect Sizes by Scenario, Averaged Over 20 Data Sets 38

Figure 3. Posterior Distribution of Effect Sizes by Scenario 39

Figure 4. MDD Study Simulation Results - Summary of Effect Sizes for HRSD-17 and PaRTS-D. At Week 4, EM Results were Not Available 45

Figure 5. Sample Size and Power Computations using a Two-Sample t-test with Two-Sided Type I Error=0.05 60

Figure 6. Scenario 1, Simulated Effect Size Ordering Using Last 20 Iterations 63

Figure 7. Scenario1, Summary of Posterior Means of Effect Size Over 20 Data Sets 64

Figure 8. Scenario1, Summary of Posterior Means of Delta Over 20 Data Sets 64

Figure 9. Scenario 2, Simulated Effect Size Ordering using Last 20 Iterations 68

Figure 10. Scenario 2, Summary of Posterior Means of Effect Size Over 20 Data Sets .. 68

Figure 11. Scenario 2, Summary of Posterior Means of Delta Over 20 Data Sets 69

Figure 12. Scenario 3, Simulated Effect Size Ordering using Last 20 Iterations 73

Figure 13. Scenario 3, Summary of Posterior Means of Effect Size Over 20 Data Sets .. 73

Figure 14. Scenario 3, Summary of Posterior Means of Delta Over 20 Data Sets 74

Figure 15. Scenario 4, Simulated Effect Size Ordering using Last 20 Iterations 78

Figure 16. Scenario 4, Summary of Posterior Means of Effect Size Over 20 Data Sets .. 78

Figure 17. Scenario 4, Summary of Posterior Means of Delta Over 20 Data Sets 79

Figure 18. Scenario 5, Simulated Effect Size Ordering using Last 20 Iterations 83

Figure 19. Scenario 5, Summary of Posterior Means of Effect Size Over 20 Data Sets .. 83

Figure 20. Scenario 5, Summary of Posterior Means of Delta Over 20 Data Sets........... 84

Figure 21. EM – Posterior Inferences on Effect Size, HRSD-17 87

Figure 22. Bayesian – Posterior Inferences on Effect Size, HRSD-17............................. 88

Figure 23. Bayesian with Strata – Posterior Inferences on Effect Size, HRSD-17 89

Figure 24. Bayesian with Strata at Week 4 – Posterior Inferences on Each Stratum on Mean Difference, HRSD-17 90

Figure 25. Bayesian with Strata at Week 6 – Posterior Inferences on Each Stratum on Mean Difference, HRSD-17 91

Figure 26. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on Effect Size, HRSD-17.. 92

Figure 27. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on Power Parameter 93

Figure 28. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on Effect Size, HRSD-17.. 94

Figure 29. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on Power Parameter 95

Figure 30. EM – Posterior Inferences on SD.. 96

Figure 31. Bayesian – Posterior Inferences on SD 97

Figure 32. Bayesian with Strata – Posterior Inferences on SD... 98

Figure 33. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on SD...... 99

Figure 34. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on SD...... 100

Figure 35. PaRTS-D Effect Size Using EM Algorithm... 101

Figure 36. PaRTS-D Effect Size Using Bayesian Approach.. 102

CHAPTER 1

INTRODUCTION

1.1 Background

The objective of clinical trials is to determine if a new medication or treatment is safe and effective. Clinical trials can have a major impact on public health. Clinical trials are conducted to find more effective ways to treat, prevent, or diagnose diseases. These research studies may also be done to find treatments with fewer side effects, or treatments that are easier for patients to tolerate. Studies evaluating efficacy and safety of new drugs, medical devices, biologics, psychological therapies, or interventions may require a significant number of subjects and extensive treatment duration. Clinical trials are major and costly undertakings for researchers. A new clinical trial examining the benefits of lowering cholesterol in women and older people may take years to complete.

Enrollment of patients in clinical trials is a continual process staggered in time. Patient recruitment in studies of chronic diseases, may take many years, so that the first endpoint for some patients can be observed when the accrual is still ongoing. For example, a depression study with a target of 300 patients, each of whom studied for six weeks, may require 12 months to enroll all 300 subjects or a year-long (per patient) bipolar trial may complete subject enrollment in 3 years. In such situations there might be ethical, practical and economic reasons for looking at the data before the planned end of the study. The design of many clinical trials includes some strategy for early stopping if an interim analysis reveals large differences between treatment groups or the possibility of no benefit from a new treatment (i.e., futility). In addition to saving time and resources, such a design feature can reduce study patients' exposure to the inferior treatment. These interim analyses can be done periodically, when a pre-specified number of participants provide required data points or at a pre-specified point in time. For instance, an interim analysis can be carried out at month 6 when half of the patients completed all 6 weeks of their participation even though this depression trial may take over 1 year to complete accrual of all patients. Interim analyses could also be carried out at equally spaced points in time or when certain information is available from the data. Statistical and clinical guidelines require that researchers document all considerations which will govern the conduct of the interim analyses in the study protocol. These considerations include, among other things, clearly stating the need for such interim analyses, the stopping rules (including any adjustment to Type I error) that will govern the conduct of the interim analyses, and the planned number of or the (information/calendar) times when such interim analyses will be carried out. Any deviations from the planned conduct of the trial could seriously flaw the validity of the study results.

Both medical ethics and the natural curiosity of researchers require ongoing monitoring of accumulating data for purposes of identifying potential study violations, assessing potential side effects, or making sure that subjects in the study are not unnecessarily given a treatment known to be inferior. This also ensures that beneficial treatments are adopted as rapidly as possible. In this dissertation, we investigate the use of the accumulated data for purposes of modifying aspects of the study design.

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. Because a study may fail to achieve its goal when the pre-study estimates or assumptions are substantially inaccurate, study designs must take this uncertainty into consideration to increase the likelihood of success. Group sequential and adaptive designs allow one to evaluate uncertainty in the planning phase without compromising the integrity of the trial. These techniques, including sample size re-estimation (blinded or unblinded), adaptive dose finding studies, and seamless phase II/III designs have been discussed extensively in the literature.

At interim points of the trial, re-evaluations of preplanned effect sizes and variance estimates may be beneficial. If the original assumptions appear erroneous, 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). The initial estimate, s_0^2 of the true variance σ^2 is typically used to compute the preplanned sample size of N_0 required to detect a clinically meaningful difference. This initial estimate is based on, among other things, previous experience with the study drugs, expert knowledge, and literature reviews. Estimates of expected mean differences share similar types of judgment calls and in many instances, these are obtained from different sets of clinical conditions.

Group sequential and adaptive designs, using unblinded interim data points, provide valid conclusions about the choice of trial design mechanisms. These are usually resource intensive procedures which require review of unblinded data at interim stages (i.e., knowledge of actual treatment assignments). Blinding and randomization are generally considered to be the most useful techniques for eliminating or minimizing bias in the design of clinical trials. Blinded experimental procedures are designed to eliminate biases produced by the anticipation of study participants, experimenters, and data analysts. In a double-blind experiment, both the participant and the experimenter are unaware of the treatment group assignment. Regulatory guidelines strongly demand that the results of interim analyses are not disseminated to the study personnel and subjects in order to preserve the integrity of the trial. To implement these procedures, independent Data Monitoring Committees need to be instituted along with data review charters to make sure that confidentiality is strictly maintained. These procedures may also inflate the Type I error rate (or equivalently, reduce power). The repeated analysis of accumulating data raises the chance of false positive findings if standard statistical methods are used at each analysis stage with no adjustments for repeated testing. The final analysis at the end must be based on the adjusted overall significance level and/or test statistics [1, 2]. It is possible to retain original power with increased sample size.

Clinical trials that demonstrate treatment differences on clinical endpoints are usually costly. In some instances, the sponsor of such a trial may have an interest in obtaining a “rough idea” of the possible magnitude of the treatment effect or even the direction of the treatment effect without unblinding while the trial is still ongoing. This information would help the sponsor to better plan future resources and new drug development strategies. Health authorities are also interested in knowing whether such assessments could impact the integrity of the trial. The following two clinical trial examples are presented to illustrate questions that can be raised prior to capturing final data points.

Depression Trial:

A large, prospective, double-blind, placebo controlled multicenter trial of adults with Major Depressive Disorder (MDD) was conducted to examine the impact of antipsychotic augmentation. The details of this trial were published by Mahmoud et al., 2007 [3]. In summary, the study included a 4-week open-label run-in phase and a 6-week double-blind placebo-controlled phase. The patients who had insufficient response to antidepressant monotherapy at the end of the open-label period were randomly allocated (1:1) to receive either active drug or matching placebo in a 6-week double-blind placebo-controlled phase in addition to their ongoing medication. Trained study personnel administered the Hamilton Rating Scale for Depression (HRSD-17 [also called HAM-D-17]) and the Clinical Global Impressions–Severity of illness instruments at each study visit. Patients completed the validated Quality of Life Enjoyment and Satisfaction Questionnaire, and the Patient-Reported Troubling Symptoms of Depression (PaRTS-D) instrument through a touch-tone telephone interactive voice response system at baseline and at each week of the double-blind period. The primary efficacy evaluation was based on the Week 4 HRSD-17 score although this study was 6 weeks long.

Assumptions for the original study design regarding expected clinical difference, population variance, drop-out rates and other clinical trial characteristics may have been based on limited experience with this drug for treatment resistant depressed patients. This is usually the case during the planning phase of a new clinical trial. Due to limited experience with the study drug, there is a degree of uncertainty during the design stage to establish the Week 4 HRSD-17 score as the primary efficacy endpoint. It is also important to monitor accumulating data to examine potential side effects, identify study violations, and examine original study design assumptions. Specifically, prior to the database lock (i.e., the observation of final data points for last enrolled patient based on required guidelines), the sponsor of this trial has the option to use information from already accrued patients to re-specify sample size or other aspects of the study design. In this example, the sponsor might have looked for guidance on the appropriateness of using treatment difference at Week 6, instead of Week 4, as the primary time point, or on the appropriateness of the performance of the patient-rated PaRTS-D score, instead of the clinician-rated HRSD-17 score, as the primary end point.

Bipolar Trial:

This randomized, double-blind, placebo controlled, international study was conducted from May 2004 to February 2007 at 32 psychiatric centers in the United States and India to evaluate adjunctive maintenance treatment with risperidone long-acting injectable (RLAI) antipsychotics in patients with bipolar disorder. The details of this trial were published by Macfadden et al., 2009 [4]. This study assessed whether adjunctive maintenance treatment with long-acting antipsychotic therapy, added to treatment as usual (TAU), delays relapse in patients with bipolar I disorder. This study included patients with bipolar I disorder with ≥4 mood episodes in the past 12 months. Following a 16-week, open-label stabilization phase with active treatment RLAI plus TAU, remitted patients entered a double-blind, placebo-controlled, relapse-prevention phase for up to 52 weeks. Randomized patients continued treatment with adjunctive RLAI plus TAU or switched to adjunctive placebo injection plus TAU. The primary outcome was time to first relapse to any mood episode. Of 275 enrolled patients, 139 entered the 52 week double-blind treatment.

Prior to the database lock, the sponsor of this trial might have looked for guidance on the adequacy of 139 subjects in the double-blind period to assess relapse rates between the two treatments.

1.2 Literature Review

In both examples above it is useful to know if we could assess the treatment effect, or even get a sense of the direction of treatment effect, based on a blinded data review. Gould and Shih (GS) [5] discussed modifying the design of ongoing trials without unblinding. First, GS provided an adjusted version of the simple one-sample variance estimator. They also proposed a procedure to estimate the within-group variance for sample size re-estimation without unblinding the clinical trial data at interim stages [2–8] using the EM algorithm [9, 10]. This procedure allowed them to obtain maximum marginal likelihood estimates (MMLEs) of within-group variability. The suggested methodology can be used to estimate not only within group variance but also the mean response difference between treatment groups. The algorithm treats the treatment group identifiers as missing data. GS demonstrated that the procedure provides a reasonable and satisfactory estimate of the common standard deviation. They acknowledged that the procedure does not reliably estimate the true difference between the treatment means.

The Gould and Shih (GS) Procedure:

Observations from treatment j follow a normal distribution with mean μ_j, $j = 0,1$ and common variance σ^2. The objective is to test H_0: $\mu_0 = \mu_1$ against H_1: $\mu_0 \neq \mu_1$. Let $\Delta = \mu_0 - \mu_1$. For purposes of determining the required sample size necessary to show a difference between treatments, Δ is assumed to be $\hat{\Delta}$ and the variance is assumed to be $\hat{\sigma}^2$. If the treatment assignments were known, $\hat{\sigma}^2$ could be computed by pooling the within-group variances. Since the assignments are unknown σ^2 can be estimated in two ways: simple adjustment of the pooled sample variance based on the difference between the means presumed by H_1; and using an EM algorithm which does not depend on H_1.

1-) GS, Simple Adjustment

Suppose that an interim sample contains λn observations from treatment 1 (active) and $(1-\lambda)n$ observations from treatment 0 (placebo); n is known and the group membership weight λ is unknown. In 1:1 treatment allocation ratio, λ is assumed to be 0.5 at each interim stage. Let Y_{ij} denote the ith $(i = 1,...,n)$ observation from group j. The overall estimate of σ^2 based on the pooled sample can be computed without unblinding and computed formally as:

$$(n-1)s_p^2 = \sum_{i,j}(Y_{ij} - \bar{Y})^2 = \sum_{i,j}(Y_{ij} - \bar{Y}_j)^2 + n\lambda(1-\lambda)(\bar{Y}_0 - \bar{Y}_1)^2$$
$$= (n-2)\hat{\sigma}^2 + n\lambda(1-\lambda)(\bar{Y}_0 - \bar{Y}_1)^2$$

where $\hat{\sigma}^2$ denotes the unknown within-group estimate of σ^2. Since the interim sample is blinded, λ and the group sample means $\bar{Y}_0$ and $\bar{Y}_1$ will be unknown, as will both terms of the last expression. However, if the alternative hypothesis $H_1\colon \mu_0 - \mu_1 = \Delta$ is true and if the size of n guarantees that $\bar{Y}_0 - \bar{Y}_1$ is reasonably close to Δ, then

$n\lambda(1-\lambda)(\bar{Y}_0 - \bar{Y}_1)^2 \approx (n-1)\Lambda(1-\Lambda)\hat{\Delta}^2$.

In particular, if $\Lambda = 0.5$, then $\hat{\sigma}^2 \approx (n-1)(s_p^2 - \hat{\Delta}^2 / 4)/(n-2)$.

2-) GS, EM Algorithm

Suppose that n interim observations have been obtained from a normal distribution with variance σ^2 and mean μ_0 or μ_1 depending on whether the individual providing the observation was in treatment group 0 or 1, respectively. Treatment group assignments are independent of observed values and unknown. Let z_i $(i = 1,...,n)$ denote the treatment group membership indicator for observation i. The z_i 's are independently distributed with $P(z_i = 1) = \lambda$, the fraction of the total sample that is allocated to treatment group 1. The log-likelihood of the interim observations y_i given the treatment indicator z_i, is

$$\ell = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(z_i(y_i-\mu_1)^2 + (1-z_i)(y_i-\mu_0)^2\Big)$$

The conditional expectation of z_i given y_i is

$$\Pr(z_i = 1 \mid y_i) = \frac{\lambda f(y_i \mid z_i = 1, \mu_0, \mu_1, \sigma)}{\lambda f(y_i \mid z_i = 1, \mu_0, \mu_1, \sigma) + (1-\lambda) f(y_i \mid z_i = 0, \mu_0, \mu_1, \sigma)}$$
$$= \left\{1 + \frac{(1-\lambda)}{\lambda}\exp\left(\frac{(\mu_1-\mu_0)(\mu_1+\mu_0-2y_i)}{2\sigma^2}\right)\right\}^{-1}$$

where, $f(y_i \mid z_i, \mu_0, \mu_1, \sigma)$ is a conditional density. The details of the EM algorithm are provided in the Methods section. The initial estimates in GS are obtained in the following manner:

Let $y_{(1)} < y_{(2)} < ... < y_{(n)}$ denote the ordered interim data, and let $\tau_i = \Phi^{-1}\left(\frac{(i-0.5)}{n}\right)$, $i = 1,...,n$, where $\Phi^{-1}(\ )$ denotes the inverse of the standard normal CDF. A simple linear regression is fit by least squares to the points $\left\{(\tau_i, y_{(i)}),\ i = 1,...,n\right\}$; let b denote the slope of the fitted line, and let a denote the intercept:

$$b = \frac{\sum \tau_i y_{(i)} - n\bar{\tau}\,\bar{y}}{\sum \tau_i^2 - n\bar{\tau}^2}, \qquad a = \bar{y} - b\bar{\tau}.$$

The initial values of $\sigma,\ \mu_1,\ \mu_0$ are $\sigma^2 = b,\ \ \mu_1 = a - b/c,\ \ \mu_0 = a + b/c$, where c is some chosen constant. The choice of c influences the estimation of the means, but not the variance. Ideally, c should be equal to $\frac{2\sigma}{\mu_0 - \mu_1}$; although b estimates σ, there is no good estimate of $(\mu_0 - \mu_1)$. GS gets around this problem as described below: In most clinical trials that use a normal approximation for estimating the sample size, the inverse of the coefficient of variation usually ranges between 0.2 and 0.5 which corresponds to about 430 and 70 patients per group, respectively, for power=0.9 and one-sided Type I error of 0.05. GS suggested taking the middle value in this range, 0.35, thus converting to $c = 2(1/0.35) = 5.71$.

Friede and Kieser questioned [11] the reliability of the within group variance estimates of the Gould and Shih approach and later provided numerous alternatives [12 – 15] for the blinded sample size evaluations. Waksman [16] examined the properties of the Gould and Shih estimates for both the within group variance and the difference between group means. He showed that the apparent non-uniqueness of the Gould and Shih estimate is due to an “apparently innocuous” alteration to the EM algorithm. When this alteration is removed, the method is valid in that it produces the maximum likelihood estimate of the within-group standard deviation. Waksman also noted that the estimates of the treatment group differences are not accurate. Xing and Ganju [17] used the enrollment order of subjects and the randomization block size to estimate the within group variance. Their simulation results showed that the variability in the estimation of treatment differences, δ, is large enough to render the estimates for δ practically useless. As a consequence, even though treatment differences are being estimated, their variability is so high that there is no risk of unblinding the trial. Miller, Friede, and Kieser [18] assessed the risk of blinded inferences based on knowledge of the randomization block size leading to inadvertent unblinding of the treatment effect. They concluded that blinded tests have reasonable power to estimate treatment effects only when the true treatment difference is several times larger than the clinically important effect assumed in the sample size computations. Consequently, blinded estimation is risky only when the study is overpowered. Overpowered studies are rare in clinical practice. The number of patients exposed to treatment, the time required to obtain the data, and hence the financial cost of the trial are all kept under control when designing a clinical trial. Overpowering may happen in the case of multiple co-primary endpoints when the study sample size is chosen for one of the endpoints.

Xie, Quan, and Zhang [19] proposed three methods, two of which were based on the EM algorithm for estimating blinded treatment effect for survival endpoints in an ongoing trial. An unblinded analysis for time-to-event data at the end of study is usually based on a log-rank test and/or a Cox proportional hazards model. The authors' first approach classifies patients into treatment and placebo groups based on their values of the surrogate endpoint (e.g., blood pressure, body weight) depending on being greater or less than the median of the pooled data. Then, the Cox proportional hazards model is applied to estimate the treatment effect. The second approach is based on an exponential regression model with the post treatment surrogate as the covariate and uses an EM algorithm for estimating the model parameters. The third approach uses a simple exponential model without conditioning on the surrogate endpoint and applies an EM algorithm for estimating model parameters. Xie, Quan, and Zhang concluded that these three methods failed to provide any reliable estimates of treatment differences because of substantial variability in parameter estimates.

1.3 Objective

The 2010 draft FDA guidance on adaptive designs [20] discusses possible study design modifications such as selection and/or order of secondary endpoints in addition to sample size re-estimation. It also indicates that the risk of bias is greatly reduced or entirely absent when adaptations rely only on blinded analyses and the blinding is strictly maintained. Although the guidance warns against presenting information on potential treatment differences, it acknowledges that blinded-analysis methods are useful and these methods do not raise concern about increasing the Type I error rate. The ICH E9 guidance [21] has also discussed the utility of blinded interim analyses.

Clinical trials are a major and costly undertaking for researchers and their planning involves careful selection of the primary and secondary endpoints. Failure to consider important secondary endpoints can limit the conclusions of clinical trials. In addition to testing the primary endpoint, examination of appropriate statistical methods to test secondary endpoints requires careful consideration for controlling the Type I error rate when performing multiple statistical tests. For instance, in schizophrenia, bipolar or depression research, investigators may include scales assessing response and remission rates, functionality, quality of life, or cognition to better characterize treatment effect. Depending on the study type and its objectives a wide variety of multiplicity approaches are available for ensuring that the probability of the Type I error is controlled within acceptable bounds. Fixed-sequence gatekeeping procedures [22] are widely used because of their ease of application and interpretation. These procedures require prospective ordering of null hypotheses for secondary endpoints after defining the primary endpoint.

This prospective ordering is based on a number of untested assumptions about the endpoints. An alternative to relying on a predefined ordering is to use information that is available to order the null hypotheses based on interim effect sizes without breaking the blind and compromising the integrity of the trial.

Instead of formal interim analyses on unblinded data, we propose using information from blinded data with masked treatment assignments. The purpose of this research is to assess the feasibility of estimating the magnitude of treatment effects on various secondary endpoints in ongoing trials without breaking the treatment blind. Our primary objective is to compare the posterior ordering of the signal-to-noise ratios (effect sizes) of endpoint using available data points. This research does not assess the impact of blinded data reviews for the secondary endpoints on the actual overall Type I error rate.

In Chapter 2, we present various Bayesian methods for ordering null hypotheses using blinded data. We assume that the endpoints are normally distributed for purposes of maximizing the power of gate-keeping approaches. For simplicity, we examine clinical trials with only 2 treatment groups, each having an unknown mean and an unknown common variance. In addition, we assume the 2 groups are compared at a fixed time point. Note that these techniques are generalizable to more than 2 treatment groups and to multiple time points (i.e., the repeated measures scenario). In Chapter 3, we show simulation results demonstrating both the utility and limitation of these methods including methods which employ the EM algorithm. Throughout the remainder of this research, we use the terminology "δ" to denote the treatment effect and "σ" the error standard deviation. In this terminology $d = \delta / \sigma$ denotes the signal to noise ratio (standardized effect size). Each of the aforementioned methods uses posterior ordering of the signal to noise $d = \delta / \sigma$ ratio among the secondary endpoints to order the secondary null hypotheses. In Chapter 4, we apply our proposed method to data collected from a depression trial and compare the results to those with unblinded estimates. In Chapter 5, we provide a discussion of the scope of the proposed methods. In Chapter 6, we conclude with additional future research activities.

CHAPTER 2

METHODS

We consider a randomized two arm trial (placebo vs. experimental treatment group) with a randomization allocation ratio of 1:1. The response variable for the k th endpoint from j th treatment group on subject i is denoted by Y_{ijk} $(i=1,...,n_j,\ j=0,1,\ k=1,2,...,K)$. $Y_{10k}, Y_{20k}, ..., Y_{n_0 0k}$ and $Y_{11k}, Y_{21k}, ..., Y_{n_1 1k}$ are sequences of normally distributed independent variables from each placebo endpoint with $Y_{i0k} \sim N(\mu_{0k}, \sigma_k^2)$ and active treatment $Y_{i1k} \sim N(\mu_{1k}, \sigma_k^2)$. We use the notation y_{ik} to denote an observed outcome with unknown treatment assignment.

In equation (1), we assume a fixed effects model for the k th endpoint, where y_{ik} is the continuous response, μ_{0k} is the mean placebo response, δ_k is the magnitude of the treatment effect, z_i is the latent binary treatment assignment, and ε_{ik} is the measurement error. We suppress the dependence of the treatment assignment on k in the sequel.

$$y_{ik} = \mu_{0k} - \delta_k z_i + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0,\ \sigma_k^2)\ \ i.i.d. \qquad (1)$$

The errors within each endpoint are assumed to be independent and normally distributed. Below, we use the notation $z_i = 1$ to indicate that subject i is assigned treatment and $z_i = 0$ to denote the placebo assignment, $(i=1,...,n_j)$. In this blinded setting, treatment assignments are unknown. We assume that lower y scores are better and the average effect of the treatment is δ_k. Suppose n observations are collected while the clinical trial is ongoing. Assume that the initial sample size N_0 was computed using the design elements for the primary endpoint. Hence, $N_0 = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2 / \delta^2$, where the notation z_α donates the α th quantile of the standard normal cumulative distribution.

The signal-to-noise ratio for endpoint k is defined to be $d_k = \delta_k / \sigma_k,\ (k = 1,...,K)$. Our objective is to compare the EM and Bayesian algorithm [23, 24] induced posterior ordering of the signal-to-noise ratios of endpoints, $P(d_1 > d_2 > ... > d_k \mid y)$.

Throughout the remainder of the dissertation, y_0 and y_1 denote generic placebo and active treatment responses, respectively. The notation y without an index represents the response with unknown treatment assignment. Generic treatment assignments are denoted by z. For a two component univariate normal mixture model, it is appropriate to specify the complete likelihood function for a given response (see equation (1)) in the form:

$$p(y, z \mid \boldsymbol{\theta}) = \lambda N(y_1 \mid \boldsymbol{\theta_1}) + (1-\lambda) N(y_0 \mid \boldsymbol{\theta_0}); \qquad \lambda = \mathrm{p(z = 1)}$$

where $\boldsymbol{\theta_0} = (\mu_0, \sigma^2),\ \boldsymbol{\theta_1} = (\mu_1, \sigma^2)$ are vector of parameters and $\boldsymbol{\lambda}$ is non-negative group membership weight. The treatment group membership of observation y_i is identified with the latent variable $z_i\ (i = 1,..., n_j)$. $N(y \mid \boldsymbol{\theta})$ designates the normal pdf of y with mean and variance parameters $\boldsymbol{\theta}$. The latent variable z_i is 1 if subject i is in the treatment group; otherwise it is 0. It is also assumed that, for $i = 1,..., n_j,\quad p(z_i = 1) = \lambda$ is fixed. The parameter λ is assumed to be 0.5 since we have only two treatment groups with a 1:1 allocation ratio. Estimates arising from maximizing the marginal likelihood are the most common way to estimate parameters of interest; however, maximum marginal likelihood estimates (MMLE) often do not have closed form solutions in settings where complex distributions are assumed. The EM algorithm is a method for finding MMLE's of parameters in statistical models, where the model depends on unobserved latent variables. Below, we introduce the EM and Bayesian algorithms to order the signal-to-noise ratios for endpoints. We assume, to begin with, that the endpoints are independent of subjects; later, we generalize this to allow for dependence on subjects. In the remainder of the dissertation y denotes the vector of observations and z denotes the vector of latent treatment assignments.

2.1 EM Algorithm

The EM algorithm starts with arbitrary parameter estimates, $\hat{\boldsymbol{\theta}}^{(t=0)}$. Let t be the current iteration index. In the E-step, the conditional expectation of the log likelihood from equation (1), $E\left\{\log p(y,z\,|\,\boldsymbol{\theta})\,|\,y,\hat{\boldsymbol{\theta}}^{(t)}\right\}$, is computed given the data and current parameter values $\hat{\boldsymbol{\theta}}^{(t)}$. In the M-step, the aforementioned conditional expectation is maximized with respect to $\boldsymbol{\theta}$. This yields the new estimates for $\hat{\boldsymbol{\theta}}^{(t+1)}$ and a distribution for $z^{(t+1)}$.

The log-likelihood of the interim observations for a given endpoint follows from (1) and is:

$$\ell \propto -\frac{n_1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_i-\mu_0+\delta\right)^2 - \frac{n_0}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_0}\left(y_i-\mu_0\right)^2 \qquad (2)$$

The aforementioned conditional expectation of latent treatment assignment at EM iteration $t+1$ is

$$P(z_i^{(t+1)}=1\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})=\frac{\lambda N(y\,|\,\hat{\mu}_0^{(t)}-\hat{\delta}^{(t)},\hat{\sigma}^{2(t)})}{\lambda N(y\,|\,\hat{\mu}_0^{(t)}-\hat{\delta}^{(t)},\hat{\sigma}^{2(t)})+(1-\lambda)N(y\,|\,\hat{\mu}_0^{(t)},\hat{\sigma}^{2(t)})}.$$

We assume $\lambda=0.5$ throughout the remainder of the discussion.

In the M-step, the mean parameters μ_0, δ, and variance parameter σ^2 at iteration $(t+1)$ are estimated as follows:

$$\hat{\mu}_0^{(t+1)}=\frac{\sum_{i=1}^{n}P(z_i^{(t+1)}=0\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})y_i}{\sum_{i=1}^{n}P(z_i^{(t+1)}=0\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})},\quad \hat{\delta}^{(t+1)}=\frac{\sum_{i=1}^{n}P(z_i^{(t+1)}=1\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})(y_i-\hat{\mu}_0^{(t+1)})}{\sum_{i=1}^{n}P(z_i^{(t+1)}=1\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})},$$

$$\hat{\sigma}^{2(\mathrm{t+1})}=\frac{\sum_{i=1}^{n}P(z_i^{(t+1)}=1\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})(y_i-\hat{\mu}_1^{(t+1)})^2+P(z_i^{(t+1)}=0\,|\,y,\hat{\boldsymbol{\theta}}^{(t)})(y_i-\hat{\mu}_0^{(t+1)})^2}{n}$$

where z_i is the treatment assignment indicator. The mean parameter μ_0 and the variance parameter σ^2 are defined analogously. The effect size, d, for each endpoint is computed at each iteration after computing current mean differences and the pooled variance.

In practice numerical methods may not identify the global maximum of the marginal likelihood function. In addition, maximum marginal likelihood estimates may be sensitive to the choice of the starting values used in the EM algorithm.

2.2 Bayesian Approach

The EM algorithm is a partially non-Bayesian methodology where the final result gives a probability distribution over the latent variables and a point estimate for $\boldsymbol{\theta}$. In the Bayesian approach, the likelihood function of the mixture distribution and the prior distributions of $\boldsymbol{\theta}$, z are combined to obtain the joint posterior distribution, which is assumed to contain all the information about the unknown parameters. Gibbs sampling, which is a special case of the Metropolis–Hastings algorithm (Gelman et al. [25]), allows us to generate samples whose distribution corresponds to the posterior distribution. Both Gibbs sampling and Metropolis–Hastings algorithm are examples of Markov Chain Monte Carlo (MCMC) algorithms. Using Bayes theorem, the posterior distribution satisfies:

$$p(z,\boldsymbol{\theta} \mid y) \propto p(y \mid z,\boldsymbol{\theta})\, p(z \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

where $p(y \mid z,\boldsymbol{\theta})$ is the likelihood, $p(z \mid \boldsymbol{\theta})$ is the prior probability of latent treatment assignment, and $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$. The conditional posterior distribution of the treatment allocation is given by:

$$p(z \mid y,\boldsymbol{\theta}) \propto p(y \mid z,\boldsymbol{\theta}).$$

This reduces to $P(z_i^{(t+1)} = 1 \mid y, \hat{\boldsymbol{\theta}}^{(t)})$ for the E-step reflecting the fact that treatment group membership labels are generated independently of one another, conditional on current values of the model parameters at each iteration. A fully Bayesian model would assume that the prior distribution $p(\boldsymbol{\theta})$ takes the form $p(\boldsymbol{\theta} \mid \zeta)$ for a given hyperparameter ζ. Each secondary endpoint would be assumed to follow a distribution from an exponential family and for each endpoint, $p(\boldsymbol{\theta}_k \mid \zeta)$ would be the conjugate Bayes prior for this family. In order to reduce the sensitivity of posterior distributions to specific choices of ζ, one should consider treating ζ as unknown with a hyperprior distribution $p(\zeta)$. For purposes of simplification and without loss of accuracy we assume fixed hyperparameters and employ $p(\boldsymbol{\theta})$ for the prior distribution of $\boldsymbol{\theta}$.

Below, $\bar{y}_j$, s_j^2 denote the mean and variance of placebo/active responses from a relevant historical experiment. Hyperparameters μ_0, μ_1, and σ^2 are assumed to have the following independent prior distributions:

$$\mu_0 \sim N(\bar{y}_0, s_0^2), \qquad \mu_1 = \mu_0 - \delta \sim N(\bar{y}_1, s_1^2),$$

σ^{-2} □ $\text{Gamma}(\alpha,\beta)$ for fixed known quantities α and β.

Regardless of the structure of the prior distribution, the posterior distribution, $p(z, \mu_0, \delta, \sigma^2 \mid y)$, does not have a simple closed form. In order to make inferences about the unknown parameters, MCMC methodologies are employed to generate samples from the posterior distribution. The Gibbs sampler is used to simulate aposteriori from Model (1). The joint posterior probability distribution of these parameters satisfies:

$$\begin{aligned} &p(z,\mu_0,\delta,\sigma^2 \mid y) \propto p(y \mid z,\mu_0,\delta,\sigma^2)\,p(z)\,p(\mu_0,\delta,\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{z_i=1}(y_i-\mu_0+\delta)^2\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{z_i=0}(y_i-\mu_0)^2\right) \\ &\times \exp\left(-\frac{(\mu_0-\delta-\bar{y}_1)^2}{2s_1^2}\right)\exp\left(-\frac{(\mu_0-\bar{y}_0)^2}{2s_0^2}\right)\left(\frac{1}{\sigma^2}\right)^{\alpha-1}\exp(-\beta\sigma^2) \qquad (3) \end{aligned}$$

We assume below that σ^2 has an inverse gamma prior with shape hyperparameter $\alpha = 3$ and a small scale hyperparameter $\beta \ (\approx 0.01)$. Under these specifications, we compute the conditional posterior distributions needed to run the Gibbs sampler. The posterior conditional distributions of parameters of interest $(\mu_0,\ \delta,\ \sigma^2)$ in equation (3) are given by:

$$\mu_0 \sim N\left(\frac{\left(\sum_{i=1} \frac{y_i + z_i\delta}{\sigma^2} \right) + \frac{\bar{y}_0}{s_0^2} + \frac{\bar{y}_1 + \delta}{s_1^2}}{\frac{n}{\sigma^2} + \frac{1}{s_0^2} + \frac{1}{s_1^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{s_0^2} + \frac{1}{s_1^2}} \right),$$

$$\delta \sim N\left(\frac{\frac{-\sum_{i=1} (y_i - \mu_0) z_i}{\sigma^2} + \frac{\mu_0 - \bar{y}_1}{s_1^2}}{\frac{\sum_{i=1} z_i}{\sigma^2} + \frac{1}{s_1^2}}, \frac{1}{\frac{\sum_{i=1} z_i}{\sigma^2} + \frac{1}{s_1^2}} \right),$$

$$\sigma^2 \sim \frac{\sum_{i=1} (y_i - \mu_0 + z_i\delta)^2 + \beta}{\chi^2_{n+3}}.$$

2.3 Bayesian Approach with Covariate

Estimates in equation (3) can be further enhanced by the inclusion of informative covariates. We assume a trial with stratified randomization where explanation of treatment differences can be further supported using a covariate (i.e., stratification factor). The full sample is divided into subsamples on the basis of a covariate (stratification factor), so that the subsamples are more homogenous. Completely randomized experiments can then be conducted within each of the subsamples. For example, one may divide the sample into subsamples by additional medication usage (e.g., two levels with *yes* and *no* groups). The joint posterior distribution in equation (3) can be extended to include estimates for stratification level parameters: δ_{yes} and δ_{no}, the treatment differences in yes and no groups, respectively. In addition to likelihoods, each group will include priors in equation (3).

Let g_i denote the known strata membership of the subject i. For instance, we set $g_i = 1$ if observation i belongs to the *yes* strata and 0 if it does not. The joint posterior distribution then takes the form:

$$p(z, \mu_0, \delta_{yes}, \delta_{no}, \sigma^2 \mid y) \propto \left(\frac{1}{\sigma^2}\right)^{2n} \exp\left(-\frac{1}{2\sigma^2} \sum_{z_i=1} \sum_{g_i=0,1} (y_i - \mu_0 + \delta_{g_i})^2\right)$$

$$\times \exp\left(-\frac{1}{2\sigma^2} \sum_{z_i=0} (y_i - \mu_0)^2\right) \exp\left(-\sum_{g_i=0,1} \frac{\left(\mu_0 - \delta_{g_i} - \bar{y}_{1g_i}\right)^2}{2s_1^2}\right) \exp\left(-\frac{\left(\mu_0 - \bar{y}_0\right)^2}{2s_0^2}\right)$$

$$\times \left(\frac{1}{\sigma^2}\right)^{\alpha-1} \exp(-\beta\sigma^2) \qquad (4)$$

The probability of missing treatment group membership for each stratum is proportional to the likelihood of its being treated. This conditional posterior probability treatment assignment at each iteration is given by:

$$P(z_i = 1) = \begin{cases} \dfrac{\exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0 + \delta_{yes}\right)^2\right)}{\exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0 + \delta_{yes}\right)^2\right) + \exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0\right)^2\right)}, \text{ Yes strata} \\ \\ \dfrac{\exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0 + \delta_{no}\right)^2\right)}{\exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0 + \delta_{no}\right)^2\right) + \exp\left(-\dfrac{1}{2\sigma^2}\sum\left(y_i - \mu_0\right)^2\right)}, \text{ No strata} \end{cases}$$

The posterior distribution for parameters of interest in equation (4) are computed to be:

$$\mu_0 \sim N\left(\frac{\left(\sum_{i=1} \dfrac{y_i + z_i(\delta_{yes} + \delta_{no})}{\sigma^2}\right) + \dfrac{\bar{y}_0}{s_0^2} + \dfrac{(\bar{y}_{1,yes} + \delta_{yes}) + (\bar{y}_{1,no} + \delta_{no})}{s_1^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{s_0^2} + \dfrac{2}{s_1^2}}, \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{s_0^2} + \dfrac{2}{s_1^2}} \right),$$

$$\delta_{yes} \sim N\left(\frac{\dfrac{-\sum\limits_{i=1,g=1} \left(y_i - \mu_0\right) z_i}{\sigma^2} + \dfrac{\mu_0 - \bar{y}_{1,yes}}{s_1^2}}{\dfrac{\sum\limits_{i=1,g=1} z_i}{\sigma^2} + \dfrac{1}{s_1^2}}, \frac{1}{\dfrac{\sum\limits_{i=1,g=1} z_i}{\sigma^2} + \dfrac{1}{s_1^2}} \right),$$

$$\delta_{no} \sim N\left(\frac{\dfrac{-\sum\limits_{i=1,g=2} \left(y_i - \mu_0\right) z_i}{\sigma^2} + \dfrac{\mu_0 - \bar{y}_{1,no}}{s_1^2}}{\dfrac{\sum\limits_{i=1,g=2} z_i}{\sigma^2} + \dfrac{1}{s_1^2}}, \frac{1}{\dfrac{\sum\limits_{i=1,g=2} z_i}{\sigma^2} + \dfrac{1}{s_1^2}} \right),$$

$$\sigma^2 \sim \frac{\sum\limits_{i=1,g=1} \left(y_i - \mu_0 + z_i\delta_{yes}\right)^2 + \sum\limits_{i=1,g=2} \left(y_i - \mu_0 + z_i\delta_{no}\right)^2 + \beta}{\chi^2_{n+3}}.$$

2.4 Bayesian Approach with Covariate and a Power Prior

Equation (4) can be further enhanced using power priors. The power prior approach provides a useful class of informative priors for Bayesian inference. The basic idea is to use the power parameter $\alpha_0 \in [0,1]$ to control the influence of the historical data on the current study [26-28]. We denote the historical data by D_0. The power prior is proportional to the product of a discounted version of the likelihood of the historical data and a prior for the size α_0 of the discount. We employ the terminology, $p(\theta)$ for the parameter prior and $p(\alpha_0)$ for the power prior of the size of the discount. We subsequently refer to the parameter α_0 as the power parameter. The contributing likelihoods take the form:

- $p(\theta \mid D_0, \alpha_0) \propto L(D_0 \mid \theta)^{\alpha_0} p(\theta) p(\alpha_0)$ is the posterior given past likelihood,
- $L(D_0 \mid \theta)$ is the past data likelihood, and
- $L(D \mid \theta)$ is the current data likelihood.

The full likelihood is proportional to $p(\theta \mid D_0, \alpha_0) L(D \mid \theta)$. Choosing a prior for the power parameter α_0 serves to characterize its likely values. The case $\alpha_0 = 0$ means that no historical data should be used; while $\alpha_0 = 1$ gives equal weight to the past data likelihood, $L(D_0 \mid \theta)$, and the likelihood of the current study $L(D \mid \theta)$, resulting in the full incorporation of the historical data. The joint posterior distribution in equation (4) can be extended to include the prior $\mathrm{Beta}(a,b)$ for α_0. The result takes the form:

$$
\begin{aligned}
p(z,\mu_0,\delta_y,\delta_n,\sigma^2,\alpha_0 \mid y) \propto & \left(\frac{1}{\sigma^2}\right)^{2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{z_i=1}\sum_{g}(y_i-\mu_0+\delta_g)^2\right) \\
& \times \exp\left(-\frac{1}{2\sigma^2}\sum_{z_i=0}(y_i-\mu_0)^2\right)\left(\frac{\sqrt{n_{1p,yes}\alpha_0}}{s_1}\right)\exp\left(-\frac{\left(\mu_0-\delta_{yes}-\bar{y}_{1,yes}\right)^2 n_{1p,yes}\alpha_0}{2s_1^2}\right) \\
& \times\left(\frac{\sqrt{n_{1p,no}\alpha_0}}{s_1}\right)\exp\left(-\frac{\left(\mu_0-\delta_{no}-\bar{y}_{1,no}\right)^2 n_{1p,no}\alpha_0}{2s_1^2}\right) \\
& \times\left(\frac{\sqrt{n_{0p}\alpha_0}}{s_0}\right)\exp\left(-\frac{\left(\mu_0-\bar{y}_0\right)^2 n_{0p}\alpha_0}{2s_0^2}\right)\left(\frac{1}{\sigma^2}\right)^{\alpha-1}\exp(-\beta\sigma^2)\alpha_0^{a-1}\left(1-\alpha_0\right)^{b-1} \qquad (5)
\end{aligned}
$$

where, n_{1p} and n_{0p} represent the number of subjects in the previous iteration that were assigned to the treatment and placebo groups, respectively. The posterior distribution of the parameters of interest in equation (5) takes the form:

$$
\begin{aligned}
\alpha_0 \propto & \ \alpha_0^{a+\frac{1}{2}}\left(1-\alpha_0\right)^{b-1} \\
& \times\exp\left(-\frac{\alpha_0}{2}\left[\left(\frac{\left(\mu_0-\bar{y}_0\right)^2 n_{op}}{s_0^2}\right)+\left(\frac{\left(\mu_0-\delta_{yes}-\bar{y}_{1,yes}\right)^2 n_{1p,yes}}{s_1^2}\right)\right]\right) \\
& \times\exp\left(-\frac{\alpha_0}{2}\left[\frac{\left(\mu_0-\delta_{no}-\bar{y}_{1,no}\right)^2 n_{1p,no}}{s_1^2}\right]\right)
\end{aligned}
$$

Although the posterior distribution of α_0 is difficult to evaluate, it is easy to update the posterior distribution using standard Monte Carlo simulation methods.

$$\mu_0 \sim N\left(\text{Mean of } \mu_0, \frac{1}{\frac{n}{\sigma^2}+\frac{n_{0p}\alpha_0}{s_0^2}+\frac{n_{1p}\alpha_0}{s_1^2}} \right);$$

where Mean of μ_0 is listed as:

$$\frac{\sum_{i=1} \frac{y_i + z_i(\delta_{yes}+\delta_{no})}{\sigma^2} + \frac{\bar{y}_0 n_{0p}\alpha_0}{s_0^2} + \frac{(\bar{y}_{1,yes}+\delta_{yes})n_{1p,yes}\alpha_0 + (\bar{y}_{1,no}+\delta_{no})n_{1p,no}\alpha_0}{s_1^2}}{\frac{n}{\sigma^2}+\frac{n_{0p}\alpha_0}{s_0^2}+\frac{n_{1p}\alpha_0}{s_1^2}},$$

$$\delta_{yes} \sim N\left(\frac{\frac{-\sum_{i=1,g=1}\left(y_i-\mu_0\right)z_i}{\sigma^2} + \frac{\left(\mu_0-\bar{y}_{1,yes}\right)n_{1p,yes}\alpha_0}{s_1^2}}{\frac{\sum_{i=1,g=1} z_i}{\sigma^2} + \frac{n_{1p,yes}\alpha_0}{s_1^2}}, \frac{1}{\frac{\sum_{i=1,g=1} z_i}{\sigma^2} + \frac{n_{1p,yes}\alpha_0}{s_1^2}} \right),$$

$$\delta_{no} \sim N\left(\frac{\frac{-\sum_{i=1,g=2}\left(y_i-\mu_0\right)z_i}{\sigma^2} + \frac{\left(\mu_0-\bar{y}_{1,no}\right)n_{1p,no}\alpha_0}{s_1^2}}{\frac{\sum_{i=1,g=2} z_i}{\sigma^2} + \frac{n_{1p,no}\alpha_0}{s_1^2}}, \frac{1}{\frac{\sum_{i=1,g=2} z_i}{\sigma^2} + \frac{n_{1p,no}\alpha_0}{s_1^2}} \right),$$

$$\sigma^2 \sim \frac{\sum_{i=1,g=1}\left(y_i-\mu_0+z_i\delta_{yes}\right)^2 + \sum_{i=1,g=2}\left(y_i-\mu_0+z_i\delta_{no}\right)^2 + \beta}{\chi^2_{n+3}}.$$

2.5 Model Selection

It is important to compare the performance of EM based methods with those employing the full posterior distribution of the parameters. By contrast with frequentist analogues, Bayesian model comparison distinguishes which likelihood and prior combinations better fit the data. These measures of performance can be used to choose a single “best” model or improve estimation via model averaging, in which expected values obtained from different models are weighted by their corresponding posterior probabilities. The deviance information criterion (DIC) provides a natural measure of performance in this setting (Spiegelhalter et al. [29]). Models with smaller DIC should be preferred to models with larger DIC.

DIC is defined as: $DIC = p_d + \bar{D}$, where p_d is a measure of model complexity and $\bar{D}$ is a Bayesian measure of how well the model fits the data. Deviance is defined as $D(\boldsymbol{\theta}) = -2\log(p(y\,|\,\boldsymbol{\theta})) + c$, where c is a constant which cancels out in calculations. The expectation of deviance, $\bar{D} = E_{\boldsymbol{\theta}}\left[D(\boldsymbol{\theta})\right]$, is the average of the log likelihood values calculated from the parameters in each sample from the posterior. The deviance evaluated at the posterior expectation is denoted by $D(\bar{\boldsymbol{\theta}})$. The effective number of parameters in the model is computed as $p_d = \bar{D} - D(\bar{\boldsymbol{\theta}})$. It follows that $DIC = 2\bar{D} - D(\bar{\boldsymbol{\theta}})$. DIC improves on BIC (Bayesian Information Criterion) and AIC (Akaike’s Information Criterion) by being sensitive to the posterior dependence between parameters. In settings like those introduced above, the DIC takes the form: $DIC = 2D(\mu_0, \delta) - D(\hat{\mu}_0, \hat{\delta})$, where

$$D(\mu_0,\delta) = E_{\mu_0,\delta}\left[\sum_{i=1} \frac{\left(X_i - \mu_0 + z_i\delta\right)^2}{\hat{\sigma}^2}\right] \qquad (6)$$

$$D(\hat{\mu}_0,\hat{\delta}) = E_{z_i}\left[\sum_{i=1} \frac{\left(X_i - \hat{\mu}_0 + z_i\hat{\delta}\right)^2}{\hat{\sigma}^2}\right] \qquad (7)$$

We employ MCMC methods to evaluate equations (6) and (7) above. The posterior expectation in equation (6) is computed by averaging over the values obtained for the expression inside the equation (6) while quantities in equation (7) are computed using MCMC averages.

2.6 MCMC: Gibbs Sampling

The Gibbs sampler is the most commonly used approach in mixture estimation (Diebolt and Robert [30]). An important feature of the Gibbs sampler is that each simulated posterior parameter value is always accepted (see equations (3), (4), and (5)). The main drawback of the Gibbs sampler in this setting is the lack of mixing when the posterior distribution has multiple isolated modes. In this situation, the inference made from the Gibbs sampler will be very poor since the allocation switch does not occur. Celeux et al. [31] extensively studied computational and inferential difficulties with mixture distributions. The Gibbs sampler cannot jump between equivalent modes of the target distribution. To circumvent this problem, we introduced a tempering scheme to switch treatment group assignments (i.e., switching labels). We employed a Metropolis–Hastings algorithm with an acceptance-rejection sampling. This involved subsampling 10% of the treatment and placebo group observations and proposing a label switch. The subsampling frequency of 10% was chosen for computational purposes. The following steps are followed, using methods developed by Tierney [32] and Chip, Greenberg [33]:

1) Draw random candidate samples from y_i for switching treatment

2) Generate u from the Uniform (0,1) distribution.

3) If $u \le \frac{\prod(1-(p(z_i=1))_i}{\prod p(z_i=1)_i}$, make label change; otherwise return to step 1.

This approach is appealing in that it encourages moves between the different modes, and is also to some extent independent of the underlying model

CHAPTER 3

SIMULATIONS

Our objective is to apply the methods discussed above to various clinical cases. The true effect sizes are assumed to be known for each endpoint. We want to investigate whether or not the researchers can obtain reliable estimates of effect sizes without knowing treatment assignments. Posterior estimates of parameters for each endpoint were examined. We infer posterior rankings of these endpoints. The performance of each method was evaluated.

For each simulated clinical trial, we assumed two treatment arms with a 1:1 treatment allocation ratio, $k = 3$ uncorrelated endpoints, y_1, y_2, and y_3, and no missing data. During the computations of blinded treatment differences, variances, and corresponding effect sizes, we mask true treatment assignments for each subject. Hence there is no way of knowing whether or not the subject is in the active or placebo group. There is also no mock treatment group identifier, such as Group A or Group B. Therefore, treatment arm assignments are missing. It is assumed that the endpoints are evaluated at one time point. Multivariate settings with autoregressive correlation structures are also assessed (see Chapter 6).

We considered 5 simulation scenarios for the variables y_1, y_2, and y_3. Each simulation scenario gave rise to parameters describing treatment means and standard deviations. These parameters were assumed to have Gaussian and inverse Gamma priors. The priors have hyper means satisfying constraints leading to various fixed effect sizes. We distinguish between parametric effect sizes, denoted below by d_1^*, d_2^*, d_3^* and the ground truth (i.e., input) fixed effect size values d_1 , d_2, d_3 . The parametric effect sizes have an empirical (posterior) distribution arising from simulations and our model assumptions. We calculate the empirical distribution under the assumption of an EM model and compare these results to the posterior distribution under the assumption of a Bayesian model. Below, we calculate the empirical (posterior) probability that the parametric effect sizes are ordered in various ways. The closer the posterior probability that the parametric effect sizes are ordered in the same way as the ground truth ordering of the effect sizes, the better the fit of the model.

3.1 Simulation Set-up

We simulated 20 data sets for each of the following scenarios. N denotes the total sample size per data set.

Scenario	d_1		d_2		d_3	N
1	0.5	=	0.5	=	0.5	170
2	0.8	>	0.5	>	0.3	170
3	1.0	>	0.5	>	0	170
4	1.0	>	0.5	>	0	72
5	1.0	>	0.5	>	0	468

The first scenario emulates the null case among the variables. The second and third scenarios provide increased variation between the effect sizes. In scenario 3, the effect sizes are further from 0.5. Scenarios 4 and 5 represent studies with different sample sizes.

In each of the 20 data sets, we considered a total sample size of 170. The sample size of 170 was chosen to ensure 90% power to detect an effect size of 0.5; two-sample t-tests were used for this purpose. In scenarios 4 and 5, total sample sizes of 468 and 72 were respectively chosen to detect effect sizes of 0.3 and 0.8. When the true effect size is 0.5, the total sample size of 468 represents an over-powered trial and the 72 subjects represents an under-powered clinical trial.

For each of the k $(k = 1, 2, 3)$ endpoints, we assumed a different prior. We used the notation $\mu_0(k,l)$ for the placebo mean parameter associated with prior k $(k = 1, 2, 3)$ and data set number l $(l = 1, ..., 20)$; similarly, $\delta(k,l)$ and $\sigma^2(k,l)$ denote the parameters associated with mean difference and variance, respectively.

We generated data sets using the following model:

$$\mu_0(k,l) \sim N(\bar{Y}_{0k}, s_{0k}^2); \quad \mu_1(k,l) = \mu_0(k,l) - \delta(k,l) \sim N(\bar{Y}_{1k}, s_{1k}^2);$$
$$\sigma^2 \sim IG(\varepsilon_k, b_k), \; k = 1, 2, 3, \;\; l = 1, ..., 20.$$

This model satisfies the conditions that $E(\mu_{0k}) = \bar{Y}_{0k}$, $E(\mu_{1k}) = \bar{Y}_{1k}$, and $Var(\mu_{0k}) = s_{0k}^2$, $Var(\mu_{1k}) = s_{1k}^2$.

Without loss of generality, we assumed that $\bar{Y}_{0k} = 10$ and $s_{0k}^2 = s_{1k}^2 = 10$. $\bar{Y}_{1k}$ was adjusted to satisfy the fixed effect size constraints for $k = 1, 2, 3$.

3.2 Simulation Results

Posterior inferences are based on 1000 iterations. The first 200 iterations are considered for the burn-in period for both the EM and Bayesian algorithms. Every 4th point after the burn-in period is stored for the Bayesian simulations to reduce correlations. The objective is to compare the performance of the EM algorithm and the Bayesian approach under the five different scenarios mentioned above.

The proportions of six different variations of effect size orderings using 1000 simulations are listed in Figure 1. The effect sizes from 20 simulated data sets produced a sample of 8000 orderings. Approximate posterior probabilities are computed from these orderings.

Simulation results for 'Bayesian Approach with Covariate' and 'Bayesian Approach with Covariate and a Power Prior' are listed in detail in the Appendix B-F.

In Figure 1, there are six possible simulated effect size orderings. For Scenario 1, each simulated effect size ordering was expected to have a 16.7% chance of occurring. Only 7 out of the aforementioned 20 simulated data sets using the EM algorithm provided results for all three variables (y_1, y_2, y_3). The ordering condition, $d_1^* > d_2^* > d_3^*$, did not occur in the EM simulations; the ordering condition $d_2^* > d_3^* > d_1^*$ occurred in 41.3% of these. Bayesian methodology produced simulations closer to the ground-truth. Recall that, in Scenario 2, the ground-truth was $d_1 > d_2 > d_3$. Again, the EM algorithm did not generate the ordering condition $d_1^* > d_2^* > d_3^*$. The Bayesian algorithm simulated the ordering condition $d_1^* > d_2^* > d_3^*$ in 20.8% of the cases and d_1^* was listed as first choice in 32.0% of cases in Bayesian versus 19.3% in EM algorithm. In Scenario 3, the EM algorithm simulated the ordering condition $d_1^* > d_2^* > d_3^*$ in 15.8% of the cases; the Bayesian algorithm produced this ordering in 26.5% of the cases. d_1^* was listed as first choice in 42.2% of cases with the Bayesian algorithm and 34.0% with the EM algorithm. The Bayesian algorithm produced the ordering condition $d_1^* > d_2^* > d_3^*$ in 19.9% and 36.2% of the cases when the sample sizes were 72 and 468, respectively. The EM algorithm produced the aforementioned ordering in 1.1% and 0% of the cases when sample sizes were 72 and 468, respectively. Again, d_1^* was listed as first choice in 33.6% and 50.7% of the cases when the sample sizes were 72 and 468 with Bayesian algorithm whereas EM listed d_1^* as a first choice in 0% and 33.6% of cases respectively.

The posterior effect size orderings showed that the EM algorithm performed poorly compared to its Bayesian counterpart. When the true ordering was $d_1 > d_2 > d_3$, the Bayesian empirical probability for identifying the ground-truth was increasing both as a function of the standardized effect size differences and sample size. The Bayesian simulations identified the true ordering less than 36.2% of the time. We note that the use of power priors in this setting increased the accuracy by up to 4%.

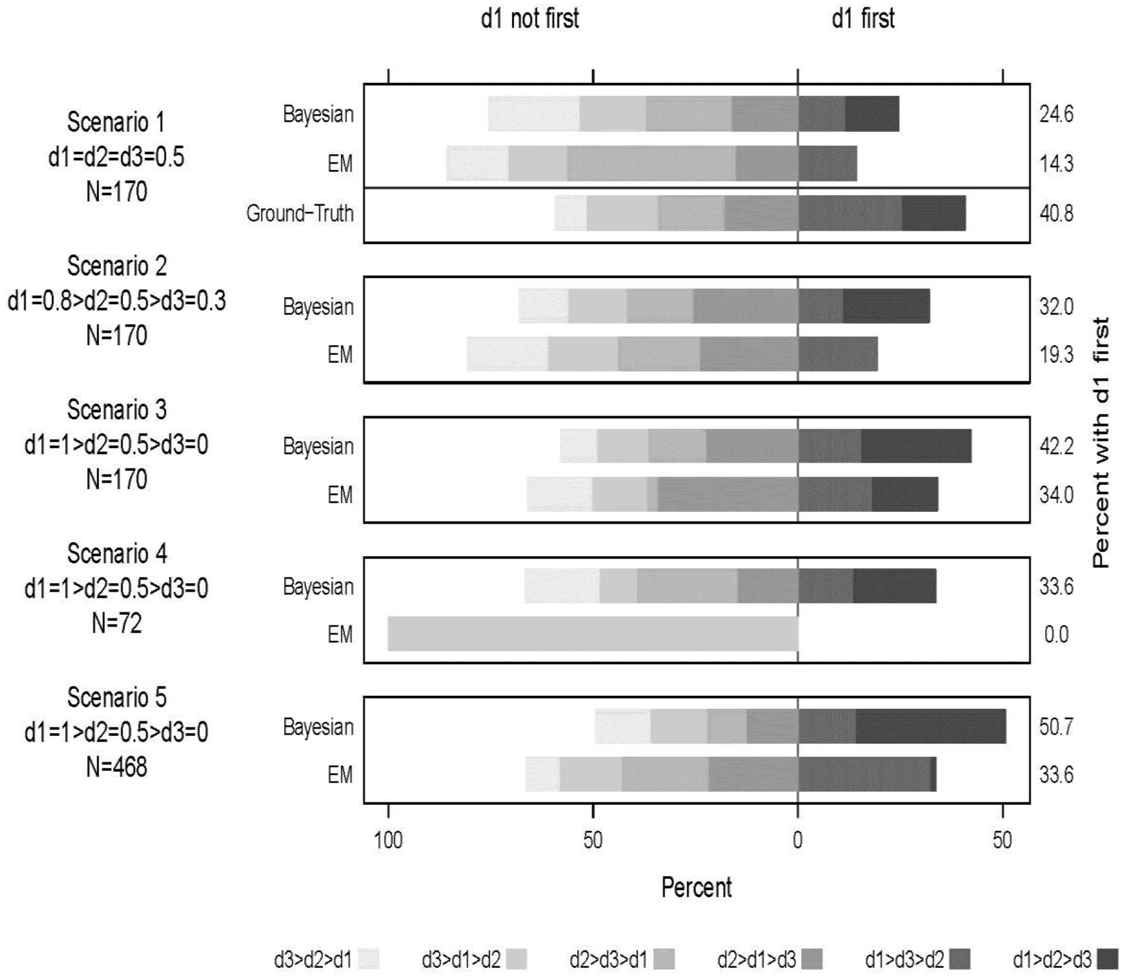


Figure 1. Posterior Probability of Effect Size Ordering by Scenario

Figures 2 and 3 show summaries of posterior inferences over 20 data sets for effect sizes in each scenario. Again, simulation results for ‘Bayesian Approach with Covariate’ and ‘Bayesian Approach with Covariate and a Power Prior’ are listed in detail in the Appendix B-F. In Figure 2, the posterior means of the parameter estimates are kept for each data set and then summarized. Figure 3 summarizes all the posterior estimates regardless of data set identifier. The EM algorithm failed to update parameter estimates for all 20 data sets. The posterior means of d^* in the Bayesian cases are overestimated compared to the true values of d . The differences obtained using the power prior estimates with starting values of 0.25 and 0.5 are negligible. The variations around estimates using the EM approach seem considerably large compared to Bayesian cases.

Posterior inferences for μ_0, σ, and δ for each endpoint were also evaluated. Both the EM and Bayesian approaches provide estimates for σ that are close to their true values. Bayesian estimates of σ had smaller posterior standard deviations as compared to their EM counterparts. Both EM and Bayesian estimates of the treatment differences, δ, overestimated their true values.

Summary tables and figures for posterior estimates by scenario are listed in the Appendix B-F. In addition, further details of the simulation set up are given in the Appendix A.

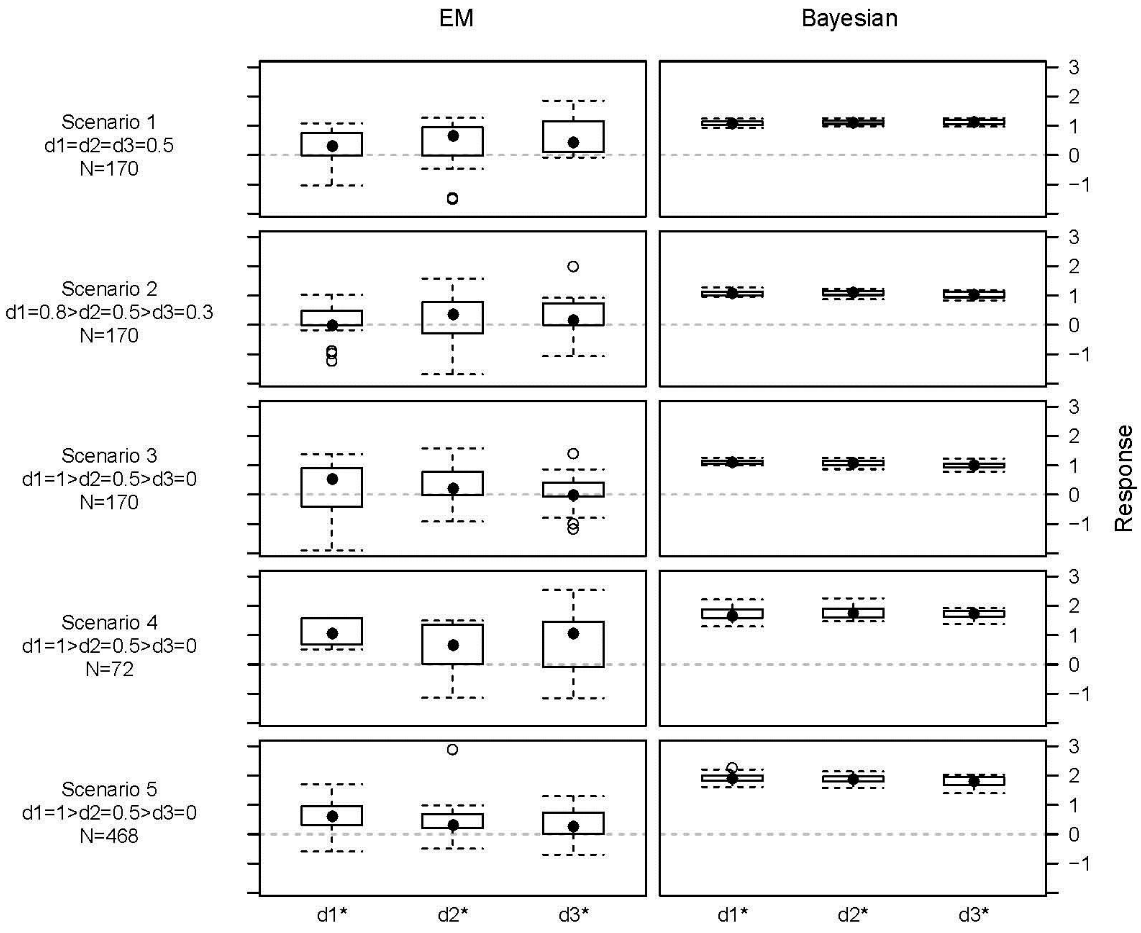


Figure 2. Posterior Distribution of Effect Sizes by Scenario, Averaged Over 20 Data Sets

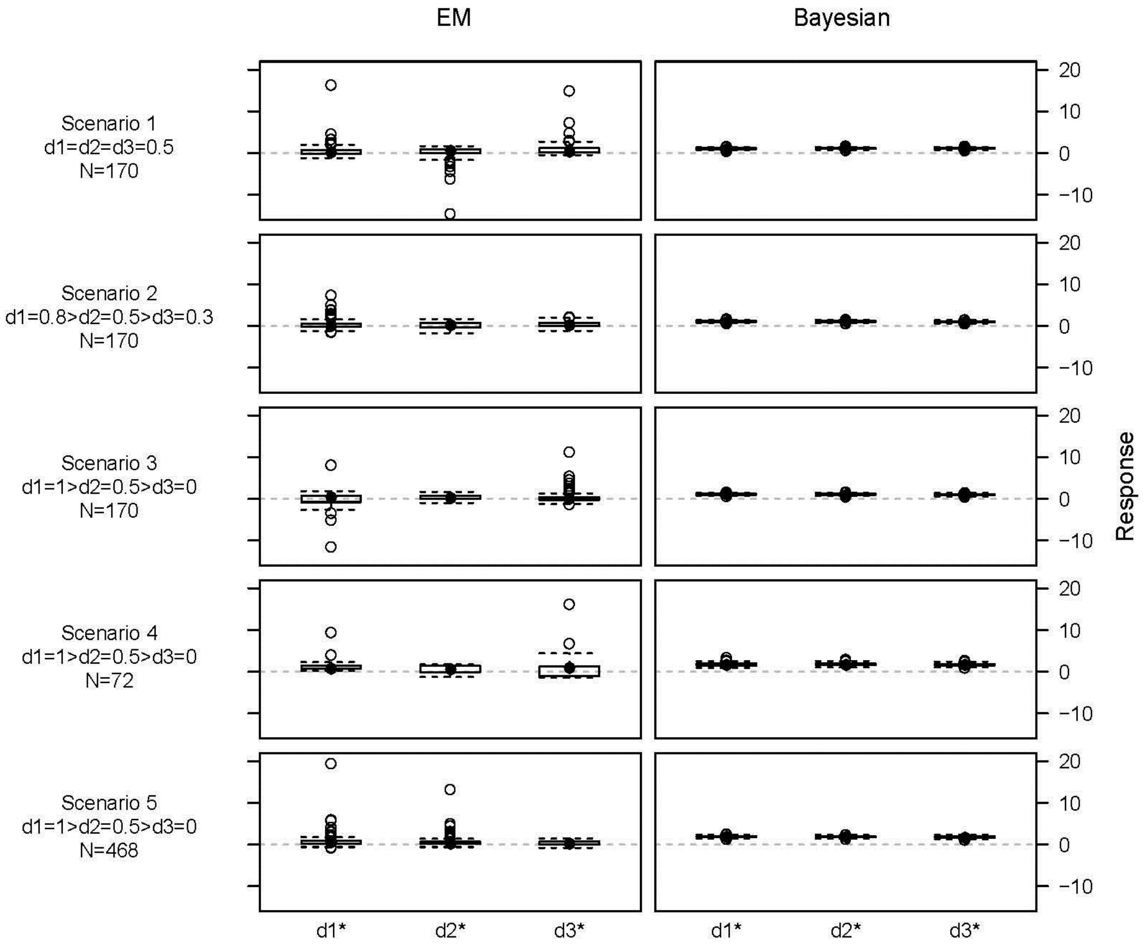


Figure 3. Posterior Distribution of Effect Sizes by Scenario

CHAPTER 4

APPLICATION

In order to apply proposed methods, we considered a placebo-controlled study in depression (Mahmoud et al. [3]). Many patients with major depressive disorder (MDD) are sub-optimally responsive to antidepressants. A variety of pharmacologic strategies are used in these patients, although relatively few have been tested systematically. Adult outpatients with MDD who had an incomplete response to ≥8 weeks of antidepressant treatment were randomly assigned to risperidone or placebo in addition to ongoing treatment for 6 weeks in a double-blind multicenter trial. Patients were randomized in a 1:1 ratio to treatment arms and were stratified by antidepressant class (SSRI or non-SSRI) and study center. The primary efficacy endpoint was the change score from baseline to Week 4 (last observation carried forward (LOCF)) in the least squares mean (LS mean ± standard error [SE]) 17-item Hamilton Rating Scale for Depression (HRSD-17) total score. At each double-blind visit (end of the open-label phase and Weeks 1, 2, 4, and 6), trained personnel administered the HRSD-17, clinician rated instrument. Scores on the HRSD-17 range from 0 to 52; higher scores indicate more severe depression. The change from baseline in HRSD-17 was analyzed at each visit using an analysis of covariance (ANCOVA) model with treatment, class of antidepressant therapy (strata), and the pooled site as factors, and baseline HRSD-17 as a covariate. Patients also independently rated their response by using the telephone interactive voice response system at baseline and weekly thereafter. The Patient-Rated Troubling Symptoms for Depression (PaRTS-D) instrument was developed to provide a more individualized assessment, from the patient's perspective, of 8 commonly reported symptoms of depression (Pandina et al. [34]). PaRTS-D scores were calculated as a percent of the maximum score possible for the four most troubling symptoms as determined at baseline.

Although this study was 6 weeks long, the primary objective of the study was evaluated at Week 4. Evaluations of blinded data while the trial was ongoing could have given researchers an opportunity to modify important design features. For example, treatment differences at Week 6 could have possibly been considered as a primary time point instead of treatment differences at Week 4. Additionally, the treatment differences using patient-rated PaRTS-D score could have possibly been used as the primary endpoint instead of the clinician-rated HRSD-17. Prior to completion of the trial, these questions were not addressed in order to preserve what was stated in the original study protocol.

The MDD study results are summarized in Table 1. The LS-Mean decrease in HRSD-17 and PaRTS-D scores from baseline to Week 4 were significantly greater with risperidone compared to placebo. Although Week 4 was the primary time point, significant between-treatment difference in HRSD-17 total and PaRTS-D scores were observed at Week 6 (95% CI at week 4 contains 0; hence not statistically significant). It is also apparent in Table 1 that week 6 effect sizes are numerically larger than those of Week 4. The 95% Confidence Intervals for least squares mean differences at week 4 for both scales overlap with the Week 6 results.

Table 1. MDD Study Results at Week 4 and Week 6 for HRSD-17 and PaRTS-D Scales Using ANCOVA Models

Parameter	Risperidone	Placebo
Hamilton Rating Scale for Depression (HRSD-17)		
Week 4 LOCF		
N	132	126
LS-Mean Change from Baseline (SE)	-8.8 (0.63)	-7.1 (0.6)
Difference on LS-Means (RIS vs. PBO) (95% CI)		-1.7 (-3.27,-0.20)
Effect Size (95% CI)		-0.3 (-0.53,-0.03)
Week 6 LOCF		
LS-Mean Change from Baseline (SE)	-10.5 (0.68)	-8.1 (0.68)
Difference on LS-Means (RIS vs. PBO) (95% CI)		-2.5 (-4.16,-0.81)
Effect Size (95% CI)		-0.4 (-0.61,-0.12)
Patient-Rated Troubling Symptoms for Depression (PaRTS-D)		
Week 4 LOCF		
N	126	120
LS-Mean Change from Baseline (SE)	-9.1 (0.88)	-7.0 (0.89)
Difference on LS-Means (RIS vs. PBO) (95% CI)		-2.1 (-4.21,0.04)
Effect Size (95% CI)		-0.3 (-0.51,0.00)
Week 6 LOCF		
LS-Mean Change from Baseline (SE)	-11.6 (0.84)	-8.1 (0.84)
Difference on LS-Means (RIS vs. PBO) (95% CI)		-3.5 (-5.57,-1.44)
Effect Size (95% CI)		-0.4 (-0.68,-0.18)

Simulation Set-up

We assumed that the trial was ongoing. Hence, there was no access to the treatment group assignments. We addressed the questions of (i) ordering null hypotheses Week 4 vs. Week 6 and (ii) ordering the effect sizes of HRSD-17 vs. PaRTS-D.

Our main objective was to examine the posterior ordering between d_1^* and d_2^* of the signal-to-noise ratios (effect sizes) of endpoints at Week 4 vs. Week 6 and/or HRSD-17 vs. PaRTS-D using the EM and Bayesian algorithms. To carry-out these computations, we assumed that the initial values for parameter estimates were based on the original sample size calculations. For the PaRTS-D evaluations, Bayesian algorithm with strata and power priors are not carried out.

We assumed the same prior distributions were utilized for all the endpoints. Recall that $\bar{y}_j$, s_j^2 $(j = 0,1)$ denote the mean and variance of placebo or active responses, respecitively, from a relevant historical experiment. For example, priors for HRSD-17 placebo and treatment means at Week 4 (H_{Wk4}) were assumed to be:

$$\mu_{04}(H_{Wk4}) \sim N(\bar{Y}_0 = 15, s_0^2 = 7^2),$$
$$\mu_{14}(H_{Wk4}) \sim N(\bar{Y}_1 = 12, s_1^2 = 7^2).$$

Two decision strategies were considered:

1- Calculate the posterior probability that the HRSD-17 effect size at Week 4 is larger than that of Week 6, $P(d^{*}(H_{Wk4}) > d^{*}(H_{Wk6}) \mid \mathbf{Y})$. If this probability is greater than 0.5, Week 4 should be considered as the primary time point.

2- Calculate 95% Credible Intervals for Week 4 and Week 6 effect sizes:

$$\underline{d}(H_{Wk4}) < d^{*}(H_{Wk4}) < \overline{d}(H_{Wk4})$$
$$\underline{d}(H_{Wk6}) < d^{*}(H_{Wk6}) < \overline{d}(H_{Wk6}).$$

If these intervals are non-overlapping, then the time point corresponding to the larger (absolute) interval should be considered as primary. If the intervals are overlapping, then the time points may not be distinguishable.

Figure 4 summarizes simulation results for HRSD-17 and PaRTS-D scores at both Weeks 4 and 6.

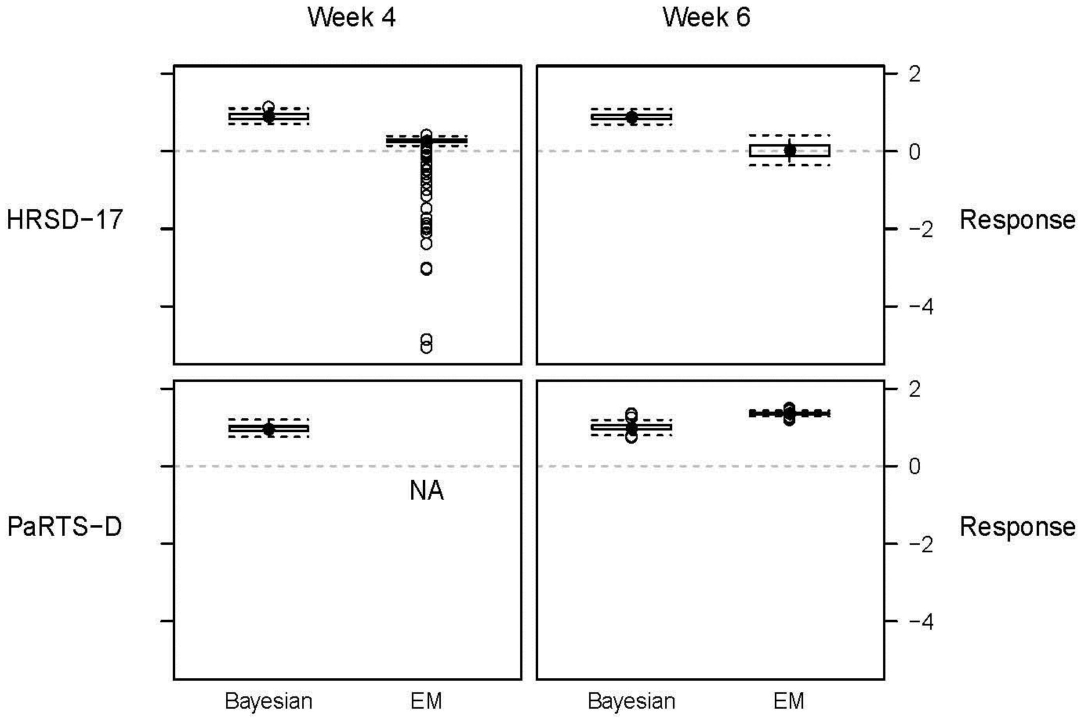


Figure 4. MDD Study Simulation Results - Summary of Effect Sizes for HRSD-17 and PaRTS-D. At Week 4, EM Results were Not Available

In Figure 4, the EM Algorithm showed that Week 4 effect sizes were numerically larger than those of Week 6 for HRSD-17 scores in contrast with the ground-truth. In both EM and Bayesian cases, the 95% credible intervals for Weeks 4 and 6 were overlapping. The Bayesian approach showed a significantly larger effect size at Week 6 for HRSD-17. The EM algorithm failed to generate results for PaRTS-D scores at Week 4 due to unbalanced latent treatment assignments. The proportions of effect size ordering for HRSD-17 were also computed, decision strategy 1. The posterior probability that the HRSD-17 effect size at Week 4 is larger than that of Week 6, $P(d^*(H_{Wk4}) > d^*(H_{Wk6}) \mid \mathbf{Y})$, was 0.477 using Bayesian approach.

The EM and Bayesian approaches provided estimates for σ that are close to the ground-truth. The EM and Bayesian estimates of the treatment differences, δ, overestimated their true values. The EM estimate of δ had a larger standard deviation compared to its Bayesian counterpart.

The DIC scores for the EM and Bayesian models are listed in Table 2 for both Weeks 4 and 6. Note that models with smaller DIC should be preferred to models with larger DIC. The Bayesian models provided smaller DIC scores.

Table 2. MDD Study Simulation Results – DIC Scores on HRSD-17 and PaRTS-D

Model	Week 4	Week 6
Hamilton Rating Scale for Depression (HRSD-17)		
EM	1.06720	1.06866
Bayesian	1.01029	1.00089
Bayesian with Strata	0.98608	1.00091
Bayesian with Strata and Power Prior 0.25	1.01086	1.00520
Bayesian with Strata and Power Prior 0.5	1.00649	1.00453
Patient-Rated Troubling Symptoms for Depression (PaRTS-D)		
EM	N/A	1.27158
Bayesian	1.00352	0.99836

CHAPTER 5

DISCUSSION AND CONCLUSION

Clinical trials are major undertakings for sponsors and investigators. To better characterize the treatment effect, a number of clinically important pre-specified secondary endpoints are needed. At the planning stage of a clinical trial, the optimal endpoints for assessing the disorder or the disease aspects that best exhibit the particular drug's effects may not be well understood. Choosing endpoints in this circumstance may be difficult at the time of study design. Obtaining data on important disease characteristics, including economic endpoints, would help sponsors, healthcare professionals, patients, and caregivers.

We investigated whether or not the researchers can obtain reliable estimates of effect sizes without knowing treatment assignments. In order to estimate treatment difference in a blinded setting, we defined a latent variable substituting for the unknown treatment assignment. We explored modifying the ordering of endpoints based on a blinded interim analysis of standardized treatment effect. Models were examined where continuous endpoints were assumed to have a normal distribution. Instead of relying on the clinician's subjective evaluations, the suggested methodologies provide numerical assistance to researchers for the purpose of ordering secondary endpoints. We simply updated the ordering of the null hypotheses based on estimating standardized treatment effects. We showed how to order secondary null hypotheses while the study is ongoing, without unblinding the treatments, without losing the validity of the testing procedure, and with maintaining the integrity of the trial.

Performance of the EM and Bayesian algorithms was evaluated using both simulations and applications. Mixture models were used for this purpose in both univariate and multivariate settings. Posterior estimates of parameters for each endpoint were examined along with posterior rankings of effect sizes based on various scenarios. The problems with initial estimates of the model parameters using the EM algorithm are well documented [11, 16 – 18]. In our simulations, we used prior distributions whose hyper parameters reflect the true values of the model parameters. In our secondary analyses, we used prior distributions whose hyper parameters reflect the original study design elements.

In order to infer posterior rankings of effect sizes, it is necessary that an algorithm generate estimates for all endpoints. The Bayesian approach generated posterior estimates for all parameters; the EM algorithm frequently failed to generate posterior parameter estimates. The Bayesian approach performed well in ordering the standardized effect sizes; the EM algorithm performed poorly. Both the Bayesian and the EM algorithms overestimated treatment differences and standardized effect sizes.

With the Bayesian algorithm, the posterior probability for identifying the ground-truth ordering increased both as a function of the effect size differences and as a function of the sample size. For large sample sizes, the proportion of times the true ordering was selected was high (above 35%) and the variability of standardized effect sizes was low.

With the EM algorithm, the probability for identifying the ground-truth ordering was low (sometimes near zero) and the variability of standardized effect sizes was high.

A high level of precision is necessary for modifying design features in blinded settings. The performance of the algorithms discussed in this paper was evaluated under the assumption that the complete set of observations was available. In practice, this is not the case. Since the standard deviations of the effect size estimates were large, it was difficult to draw proper conclusions regarding the magnitude of effect sizes. If the researcher wants to change the design of an ongoing trial, clear arguments for the changes should be provided and the arguments should be based on the clinical, functional, and economic importance of the endpoints.

Previous research about estimation of blinded treatment differences used the EM algorithm which failed to provide reliable conclusions about the treatment effect. Approaches that employ the EM algorithm to estimate treatment differences in blinded settings do not provide reliable conclusions about ordering the null hypotheses. We have demonstrated with simulation studies that Bayesian algorithms performed better than existing EM algorithm counterparts in ordering effect sizes. We have shown that the Bayesian approach provides results close to the ground-truth of simulations. In our secondary analysis, we have demonstrated using the Bayesian approach that Week 6 would have been chosen as a primary endpoint instead of Week 4.

CHAPTER 6

FUTURE RESEARCH

The current research findings do not account for within-subject correlations among endpoints nor among the measurement times. In many applications, it is of interest to assess the dependence structure (i.e., the variance-covariance structure) in multivariate longitudinal data. Identifying such dependence is challenging due to the dimensionality involved. If the dependence structure between different responses (endpoints) is not of interest, in other words, the focus is the time-varying relationship between the different longitudinal responses, then one can use multivariate normal linear models, which allow correlations between random effects in component models for each response. A natural question in Chapter 4 for the comparison of HRSD-17 scores at Week 4 and Week 6 is "why should we use a different treatment assignment given that the subject treatment assignment does not change during the trial from visit to visit." This specific question is also important for the comparisons of effect sizes among the secondary endpoints.

Researchers often have specific expectations with respect to changes over time (e.g., response increases on average over time). When the study population is divided into different subgroups, such as in Chapter 2, researchers may also be interested in the difference between these groups over time. In other words, group 1 may respond similarly to group 2 at Week 4, but not at Week 6 or vice-verse.

Multivariate Model

Using the aforementioned terminology we now present a multivariate model to explain treatment effects over time. We focus on the time-varying relationships between responses associated with a single endpoint. Multivariate treatment models serve to exploit the correlation between the observed responses of a patient at different times. One option is to estimate the matrix of correlations between the treatment effects of a patient at different times. This option has the disadvantage that the typical conjugate priors have a Wishart distribution; it is difficult to specify priors for the specific temporal correlations in this setting. In view of this, we adopt a model in which correlations are specified via specific parameters. In this model the parameters, $\phi_{t't}$ are used to specify the correlations between the treatment response at times t' and t $(t' < t)$. Below, we assume time units $t = 1$ (Week 1), $t = 2$ (Week 2), etc.. The model we adopt was first suggested in a paper by Pourahmadi and Daniels [35]; it was first employed in economic settings where it was called the "GARP" model. The model is summarized as follows:

$$y_{it} = \begin{cases} \mu_0 - \delta z_i + \varepsilon_i & \text{if } \ t=1 \\ \mu_0 - \delta z_i + \sum_{t'=1}^{t-1} \phi_{t't}\left\{y_{it'} - \mu_0 + \delta z_i\right\} + \varepsilon_{it} & \text{if } \ t=2,3,4 \end{cases}$$

$$\varepsilon_{it} \sim N(0, \sigma^2)$$

The parameters of this model are $\theta = (\mu_0, \delta, \sigma, \phi_{t't}[1 \le t' < t < 4])$. We assume a uniform prior for all of these parameters. Posterior inference for estimating θ uses a Gibbs sampling MCMC algorithm. Gibbs sampling exploits the conditional distributions of each parameter given the rest. In what follows, we outline these conditional posterior distributions. We used the notation $\theta_{-\bullet}$ to denote the entire class of θ parameters excluding $\bullet$. The latent treatment assignment, averaged μ over time points, averaged δ over time points, and σ^2 have the following form:

$$P(z_i = 1 \mid y, \theta_{-z_i}) = \frac{\exp\left\{-\frac{1}{2}\sum_{t=1}^{4} r(i,t,1)^2\right\}}{\exp\left\{-\frac{1}{2}\sum_{t=1}^{4} r(i,t,1)^2\right\} + \exp\left\{-\frac{1}{2}\sum_{t=1}^{4} r(i,t,0)^2\right\}},$$

$$r(i,t,1) = y_{it} - \mu + \delta - \sum_{t'<t} \phi_{tt'}\left\{y_{it'} - \mu + \delta\right\},$$

$$r(i,t,0) = y_{it} - \mu - \sum_{t'<t} \phi_{tt'}\left\{y_{it'} - \mu\right\}$$

$$\left(\mu \mid \theta_{-\mu}\right) \sim \mathrm{N}\left(\frac{\sum_{t=1}^{4}\left\{\sum_{i=1}^{n}\left(y_{it} + \delta z_i - \sum_{t'<t}\phi_{t't}\left(y_{it'} + \delta z_i\right)\right)\left(1 - \sum_{t'<t}\phi_{t't}\right)\right\}}{n\sum_{t=1}^{4}\left(1 - \sum_{t'<t}\phi_{t't}\right)^2}, \frac{1}{n\sum_{t=1}^{4}\left(1 - \sum_{t'<t}\phi_{t't}\right)^2}\right)$$

$$\left(\delta \mid \theta_{-\delta}\right) \sim N\left(\frac{\sum_{t=1}^{4}\left\{\sum_{i=1}^{n}\left(y_{it} - \mu - \sum_{t'<t}\phi_{t't}\left(y_{it'} - \mu\right)\right)\left(\sum_{t'<t}\phi_{t't} - 1\right)\right\}}{\sum_{i=1}^{n} z_i\left(1 - \sum_{t'<t}\phi_{t't}\right)^2}, \frac{1}{\sum_{i=1}^{n} z_i\left(1 - \sum_{t'<t}\phi_{t't}\right)^2}\right)$$

$$\sigma^2 \sim \frac{\sum_{t=1}^{4}\left\{\sum_{i=1}^{n}\left(y_{it}-\mu+\delta z_i-\sum_{t'<t}\phi_{t't}\left(y_{it'}-\mu+\delta z_i\right)\right)^2\right\}}{\chi^2_{4n+3}}$$

REFERENCES CITED

1. Jennison C, Turnbull B. *Group Sequential Methods With Applications to Clinical Trials*. Chapman and Hall/CRC; 1999.

2. Cui L, Hung MJ, Wang SJ. Modification of sample size in group sequential trials. *Biometrics* 1999; 55:853–857.

3. Mahmoud AR, Pandina JG, Turkoz I, Kosik-Gonzalez C, Canuso MC, Kujawa JM, and Gharabawi GM. Risperidone for treatment-refractory major depressive disorder. *Annals of Internal Medicine* 2007; 147: 593-602.

4. Macfadden W, Alphs L, Haskins JT, Turner N, Turkoz I, Bossie C, Kujawa M, Mahmoud R. A randomized, double-blind, placebo controlled study of maintenance treatment with adjunctive risperidone long-acting therapy in patients with bipolar I disorder who relapse frequently. *Bipolar Disord* 2009: 11: 827–839.

5. Gould AL, Shih WJ. Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* 1998; 17: 89-100.

6. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statististics Theory and Methods* 1992; 21: 2833–2853.

7. Shih WJ. Sample size reestimation in clinical trials. In *Biopharmaceutical sequential statistical applications*, Peace K (ed.). Marcel Dekker: New York, 1992; 285–301.

8. Shih WJ. Sample size reestimation for triple blind clinical trials. *Drug Information Journal* 1993; 27: 761–764.

9. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; 39: 1–38.

10. Little RJA, Rubin DB. *Statistical analysis with missing data.* Wiley: New York, 1987.

11. Friede T, Kieser M. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine* 2002; 21: 165–176.

12. Kieser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* 2000; 19: 901-911.

13. Friede T, Kieser M. A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine* 2001; 20: 2625-2643.

14. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* 2003; 22: 3571-3581.

15. Friede T, Kieser M. Blinded sample size assessment in non-inferiority and equivalence trials. *Statistics in Medicine* 2003; 22: 995-1007.

16. Waksman A J. Assessment of the Gould-Shih procedure for sample size re-estimation. *Pharmaceutical Statistics* 2007; 6: 53-65.

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

18. Miller F, Friede T, Kieser T. Blinded assessment of treatment effects utilizing information about randomization block length. *Statistics in Medicine* 2009; 28: 1690-1706.

19. Xie J, Quan H, Zhang J. Blinded assessment of treatment effects for survival endpoint in an ongoing trial. *Pharmaceutical Statistics* 2012.

20. Guidance for industry: Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf

21. ICH E9. Statistical principles for clinical trials. http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.

22. Demitrienko A., Tamhane A., Bretz F. *Multiple testing problems in pharmaceutical statistics* 2010; Chapman&Hall/CRC Biostatistics Series.

23. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov chain Monte Carlo in practice* 1996; Chapman&Hall/CRC Interdisciplinary Statistics.

24. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. *Machine Learning* 2003; 50: 5-43.

25. Gelman A, Carlin JB, Stern HS, Rubin D. *Bayesian Data Analysis* 2004; Chapman&Hall/CRC Interdisciplinary Mathematics.

26. Ibrahim J, Chen MH. Power prior distributions for regression models. *Statistical Science* 2000; 15:46-60.

27. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 2006; 17: 95-106.

28. Hobbs BP, Carlin BP, Mandrekar S, and Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 2011; 67:1047-1056.

29. Spiegelhalter DJ, Best NG, Carlin BP, Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistics Society, Series B (Statistical Methodology)* 2002; 64: 583–639.

30. Diebolt J, Robert CP. Estimation of finite mixture distributions by Bayesian Sampling. *Journal of Royal Statistics Society, Series B* 1994; 56: 363-375.

31. Celeux G, Hurn M, Robert CP. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 2000; 95:957-970.

32. Tirney, L. Markov Chains for exploring posterior distributions. *The Annals of Statistics* 1994; 22:1701-1728.

33. Chib, S. and Greenberg, E. Understanding the Metropolis–Hasting Algorithm. *The American Statistician* 1995; 49: 327-335.

34. Pandina JG, Revicki AD, Kleinman L, Turkoz I, Wu HJ, Kujawa JM, Mahmoud R, Gharabawi GM. Patient-rated troubling symptoms of depression instrument results correlate with traditional clinician- and patient-rated measures: A secondary analysis of a randomized, double-blind, placebo-controlled trial. *Journal of Affective Disorders* 2009; 118: 139–146.

35. Pourahmadi M, Daniels MJ. Dynamic Conditionally Linear Mixed Models for Longitudinal Data. *Biometrics* March 2002; 58:225-231.

APPENDICES

APPENDIX A

SIMULATION DETAILS

Scenario 1: $d_1 = d_2 = d_3 = 0.5;$ N=170

Scenario 2: $d_1 = 0.8 > d_2 = 0.5 > d_3 = 0.3;$ N=170

Scenario 3: $d_1 = 1 > d_2 = 0.5 > d_3 = 0;$ N=170

Scenario 4: $d_1 = 1 > d_2 = 0.5 > d_3 = 0;$ N=72

Scenario 5: $d_1 = 1 > d_2 = 0.5 > d_3 = 0;$ N=468

- Two arm trials with a 1:1 treatment assignment ratio are considered. For each case, three random variables are created, so that the difference between the control distribution and an experimental distribution corresponds to one of the suggested effect size cases given in the first bullet point.

- Notation: $\mu_0(k,l)$, $\delta(k,l)$, and $\sigma^2(k,l)$ are the parameters associated with prior k $(1,2,3)$ and data set number l $(1,...20)$. We always assume that $z_1,...,z_n$ are iid binary with probability 0.5 of being 0 or 1.

- Generate data sets including continuous response variables with $\mu_0,\ \delta,\ \sigma^2, z$ according to

$$\begin{aligned}
&\mu_0(1,l) \sim N(\bar{Y}_{01}, s_{01}^2); \quad \mu_1(1,l) = \mu_0(1,l) - \delta(1,l) \sim N(\bar{Y}_{11}, s_{11}^2); \quad \sigma^2 \sim IG(\varepsilon_1, b_1),\\
&\mu_0(2,l) \sim N(\bar{Y}_{02}, s_{02}^2); \quad \mu_1(2,l) = \mu_0(2,l) - \delta(2,l) \sim N(\bar{Y}_{12}, s_{12}^2); \quad \sigma^2 \sim IG(\varepsilon_2, b_2),\\
&\mu_0(3,l) \sim N(\bar{Y}_{03}, s_{03}^2); \quad \mu_1(3,l) = \mu_0(3,l) - \delta(3,l) \sim N(\bar{Y}_{13}, s_{13}^2); \quad \sigma^2 \sim IG(\varepsilon_3, b_3).
\end{aligned}$$

These three distributions above satisfy the condition for each endpoint that $E(Y_0) = \mu_0$, $E(Y_1) = \mu_1$, and $Var(Y_0) = Var(Y_1) = \sigma^2$.

- Each true input clinical trial data set is based on $\mu_0 \sim N(10,10)$ and μ_1 with normal distribution with variance 10^2 and corresponding mean value to generate effect sizes for a given case.
- Initial values of $(\mu_0,\ \delta,\ \sigma^2)$ for simulations are based on the true input data set parameters and the initial power parameter estimate α_0 is given as 0.25 and 0.5.
- Posterior inferences are based on 1000 iterations. The first 200 iterations are considered to be the burn-in period for both EM and Bayesian algorithms. Every 4th point after the burn-in period is stored for the Bayesian simulations to reduce correlations to emulate thinning process. Eight hundred simulation results in EM and 200 results in Bayesian methods are summarized.
- In addition to posterior parameter estimates, the empirical probability of each triplet combination of effect size ordering is examined. In each case for a given data set, the last 20 simulation results are kept for effect size orderings. For each of the study data sets and three effect sizes, a total of 20×20×20=8000 possible combinations are available. When this exercise is repeated for 20 different study data sets, then a total of 160,000 possible orderings are available.

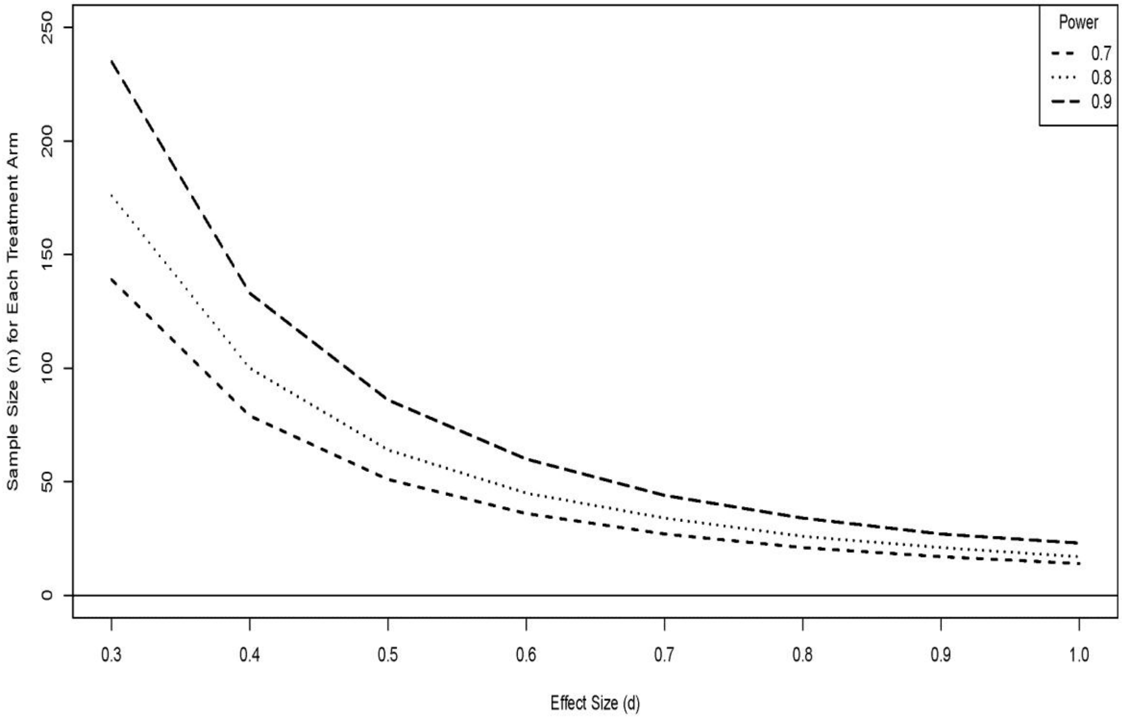


Figure 5. Sample Size and Power Computations using a Two-Sample t-test with Two-Sided Type I Error=0.05

APPENDIX B

SIMULATION RESULTS, SCENARIO 1

Table 3. Scenario 1, Input Data Set Parameter Characteristics

Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
Placebo Mean	y1	20	10.2 (0.95)	10.2	8.54; 12.32
	y2	20	10.1 (0.91)	9.9	8.68; 11.89
	y3	20	10.1 (0.91)	10.1	8.27; 12.04
Sigma (STD)	y1	20	10.2 (0.64)	10.2	9.15; 11.13
	y2	20	10.0 (0.53)	10.0	9.09; 10.94
	y3	20	10.0 (0.47)	10.0	9.30; 11.02
Delta	y1	20	5.1 (0.29)	5.0	4.63; 5.52
	y2	20	5.0 (0.27)	5.0	4.42; 5.49
	y3	20	5.0 (0.27)	5.0	4.51; 5.52
Effect Size, d	y1	20	0.50 (0.01)	0.50	0.48; 0.52
	y2	20	0.50 (0.01)	0.50	0.48; 0.52
	y3	20	0.50 (0.01)	0.50	0.48; 0.52

The proportion of six different variations of effect size orderings using input data sets are listed in Table 1. These effect sizes with 20 data sets produce a total of 8000 possibilities. Each ordering condition is expected to have a 16.7% chance of occurring in given that each condition is equally likely.

Table 4. Scenario 1, Effect Size Ordering of Input Data Sets

Order	Condition	Count	Percent (%)
1	d1>d2>d3	1212	15.2
2	d1>d3>d2	2053	25.7
3	d2>d1>d3	1455	18.2
4	d2>d3>d1	1293	16.2
5	d3>d1>d2	1395	17.4
6	d3>d2>d1	592	7.4

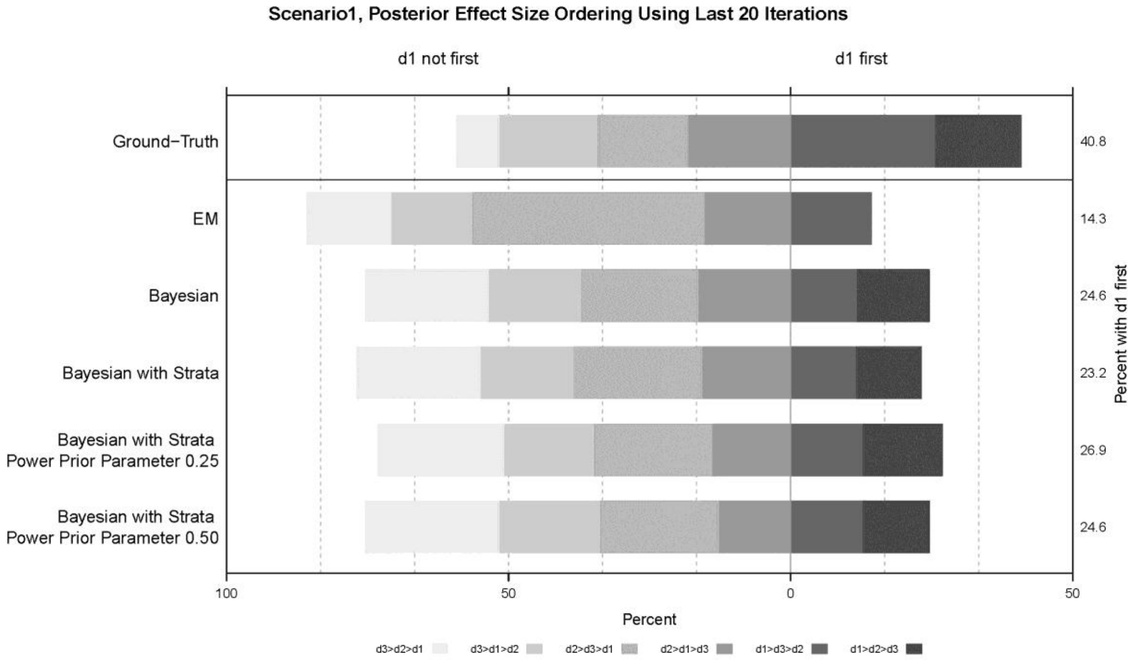


Figure 6. Scenario 1, Simulated Effect Size Ordering Using Last 20 Iterations

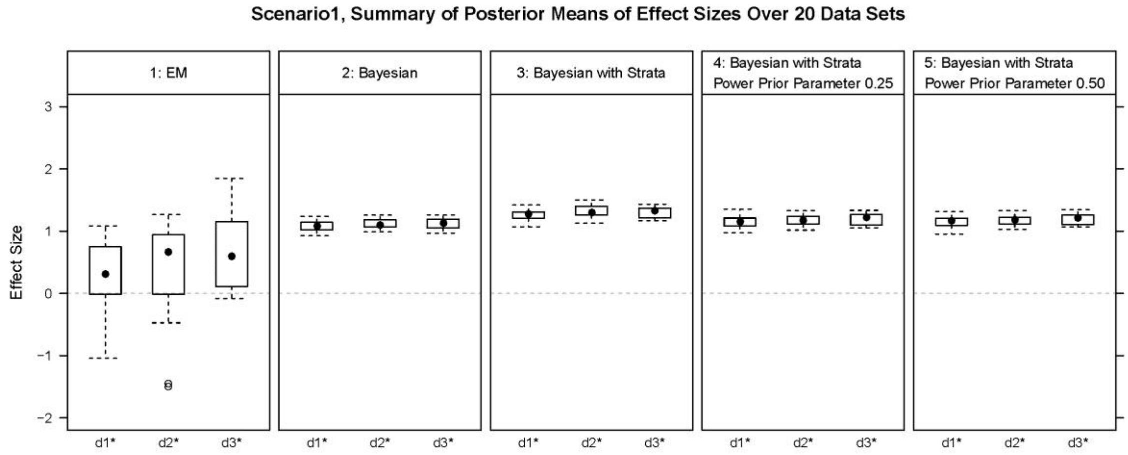


Figure 7. Scenario1, Summary of Posterior Means of Effect Size Over 20 Data Sets

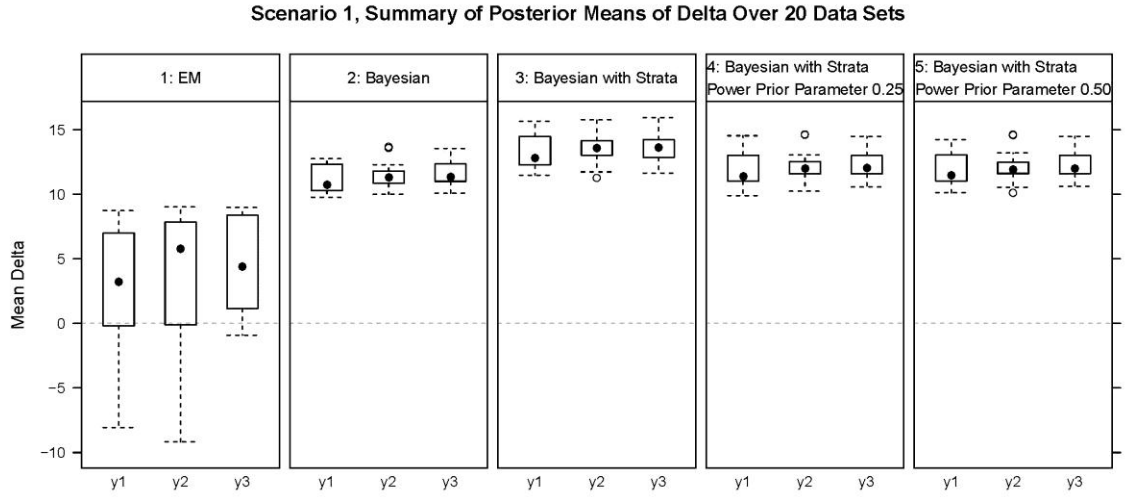


Figure 8. Scenario1, Summary of Posterior Means of Delta Over 20 Data Sets

Table 5. Scenario 1, Summary of Posterior Means of Parameters Over 20 Data Sets

Analysis Type		Statistics			
Parameter	Variable	N	Mean (SD)	Median	95% CI
EM					
Sigma	y1	15	9.00 (2.055)	10.07	2.9 ; 10.8
	y2	18	8.49 (1.518)	8.87	6.1 ; 11.2
	y3	15	8.88 (2.016)	9.77	4.6 ; 11.2
Delta	y1	15	2.63 (4.996)	3.26	-8.1 ; 8.7
	y2	18	3.07 (5.893)	5.81	-9.2 ; 9.0
	y3	15	4.42 (3.651)	4.42	-0.9 ; 9.0
Effect Size d*	y1	15	81.45 (314.5)	0.31	-1.0 ; 1218.2
	y2	18	0.35 (0.836)	0.67	-1.5 ; 1.3
	y3	15	0.69 (0.618)	0.60	-0.1 ; 1.8
Bayesian					
Sigma	y1	20	10.44 (0.665)	10.47	9.4 ; 11.5
	y2	20	10.28 (0.490)	10.30	9.4 ; 11.2
	y3	20	10.31 (0.518)	10.27	9.6 ; 11.4
Delta	y1	20	11.19 (1.051)	10.75	9.8 ; 12.8
	y2	20	11.43 (0.984)	11.32	10.0 ; 13.7
	y3	20	11.53 (0.953)	11.36	10.1 ; 13.6
Effect Size d*	y1	20	1.08 (0.086)	1.08	0.9 ; 1.2
	y2	20	1.12 (0.078)	1.10	1.0 ; 1.3
	y3	20	1.12 (0.084)	1.13	1.0 ; 1.3

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

Table 5 Continued.

Bayesian with Strata					
Sigma	y1	20	10.62 (0.728)	10.68	9.5 ; 11.7
	y2	20	10.39 (0.497)	10.37	9.5 ; 11.3
	y3	20	10.48 (0.527)	10.47	9.7 ; 11.6
Delta	y1	20	13.30 (1.253)	12.82	11.5 ; 15.7
	y2	20	13.64 (1.137)	13.59	11.3 ; 15.8
	y3	20	13.64 (1.073)	13.66	11.6 ; 15.9
Effect Size d*	y1	20	1.26 (0.093)	1.28	1.1 ; 1.4
	y2	20	1.32 (0.098)	1.30	1.1 ; 1.5
	y3	20	1.31 (0.082)	1.34	1.2 ; 1.4
Bayesian with Strata and Power Prior 0.25					
Sigma	y1	20	10.45 (0.677)	10.53	9.4 ; 11.5
	y2	20	10.26 (0.504)	10.26	9.4 ; 11.1
	y3	20	10.33 (0.526)	10.28	9.6 ; 11.5
Delta	y1	20	11.94 (1.250)	11.42	9.9 ; 14.5
	y2	20	12.14 (1.113)	12.03	10.2 ; 14.7
	y3	20	12.29 (1.087)	12.05	10.6 ; 14.5
Effect Size d*	y1	20	1.15 (0.099)	1.16	1.0 ; 1.4
	y2	20	1.19 (0.086)	1.18	1.0 ; 1.3
	y3	20	1.20 (0.091)	1.23	1.1 ; 1.3
Bayesian with Strata and Power Prior 0.5					
Sigma	y1	20	10.47 (0.68)	10.49	9.3 ; 11.5
	y2	20	10.27 (0.52)	10.27	9.4 ; 11.2
	y3	20	10.32 (0.51)	10.34	9.6 ; 11.4
Delta	y1	20	11.97 (1.21)	11.48	10.1 ; 14.2
	y2	20	12.10 (1.11)	11.93	10.1 ; 14.6
	y3	20	12.30 (1.07)	12.01	10.6 ; 14.5
Effect Size d*	y1	20	1.15 (0.09)	1.17	1.0 ; 1.3
	y2	20	1.18 (0.09)	1.18	1.0 ; 1.3
	y3	20	1.20 (0.09)	1.22	1.1 ; 1.3

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

APPENDIX C

SIMULATION RESULTS, SCENARIO 2

Table 6. Scenario 2, Input Data Set Parameter Characteristics

		Statistics			
Parameter	Variable	N	Mean (SD)	Median	95% CI
Placebo Mean	y1	20	10.5 (0.61)	10.5	9.47; 12.02
	y2	20	10.1 (0.82)	10.3	8.54; 11.82
	y3	20	9.7 (1.09)	9.7	6.54; 11.84
Sigma (STD)	y1	20	10.2 (0.52)	10.4	9.08; 11.08
	y2	20	10.0 (0.52)	10.0	9.13; 11.29
	y3	20	9.9 (0.49)	9.8	9.17; 10.95
Delta	y1	20	8.8 (0.91)	8.8	7.09; 11.37
	y2	20	5.3 (0.68)	5.3	4.12; 6.25
	y3	20	2.3 (1.15)	2.2	-0.65; 3.77
Effect Size, d	y1	20	0.9 (0.01)	0.9	0.71; 1.09
	y2	20	0.5 (0.06)	0.5	0.42; 0.61
	y3	20	0.2 (0.11)	0.2	-0.06; 0.36

Table 7. Scenario 2, Effect Size Ordering of Input Data Sets

Order	Condition	Count	Percent (%)
1	d1>d2>d3	8000	100.0

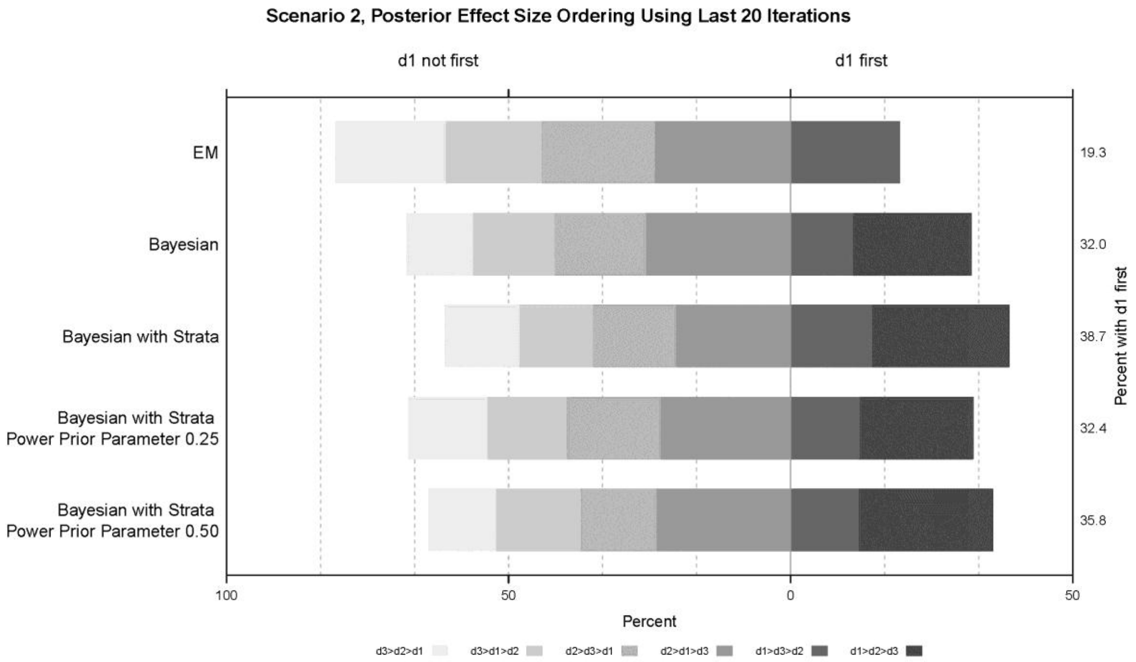


Figure 9. Scenario 2, Simulated Effect Size Ordering using Last 20 Iterations

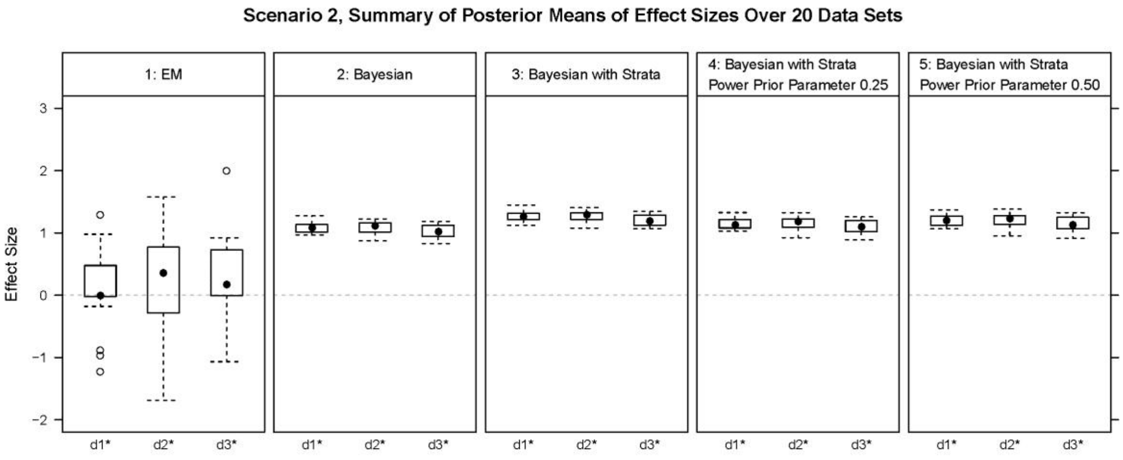


Figure 10. Scenario 2, Summary of Posterior Means of Effect Size Over 20 Data Sets

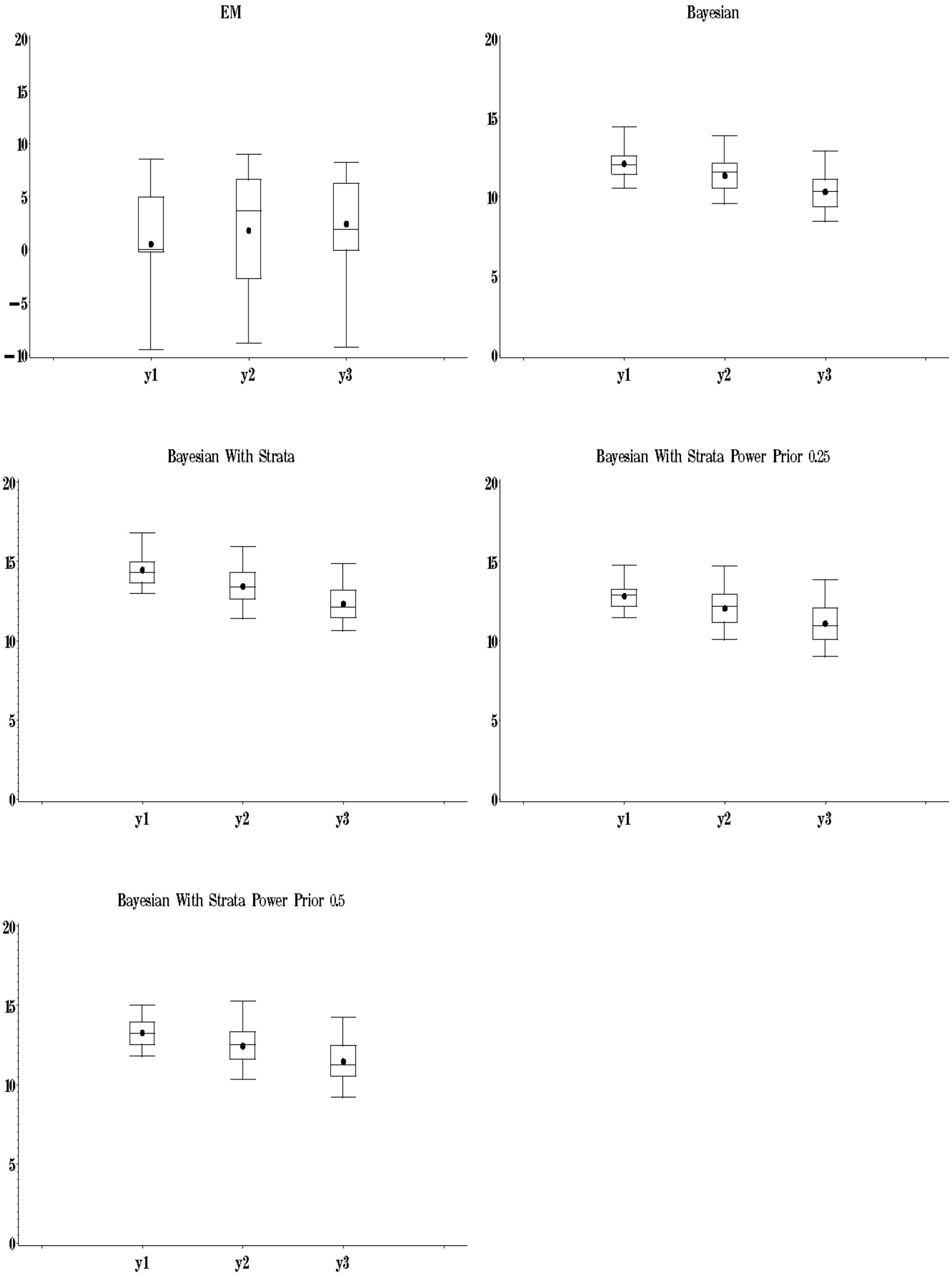


Figure 11. Scenario 2, Summary of Posterior Means of Delta Over 20 Data Sets

Table 8. Scenario 2, Summary of Posterior Means of Parameters Over 20 Data Sets

Analysis Type Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
EM					
Sigma	y1	17	9.99 (1.445)	10.55	7.8 ; 11.6
	y2	16	8.75 (1.782)	9.32	5.1 ; 10.8
	y3	18	9.07 (1.474)	9.31	3.8 ; 10.5
Delta	y1	17	0.48 (5.360)	-0.04	-9.5 ; 8.5
	y2	16	1.78 (5.961)	3.64	-8.9 ; 9.0
	y3	18	2.40 (4.413)	1.84	-9.3 ; 8.2
Effect Size d*	y1	17	0.06 (0.673)	-0.00	-1.2 ; 1.3
	y2	16	0.23 (0.862)	0.36	-1.7 ; 1.6
	y3	18	0.33 (0.638)	0.18	-1.1 ; 2.0
Bayesian					
Sigma	y1	20	11.18 (0.508)	11.18	10.2 ; 11.9
	y2	20	10.43 (0.582)	10.32	9.5 ; 11.7
	y3	20	10.08 (0.479)	10.02	9.4 ; 11.0
Delta	y1	20	12.07 (0.890)	11.97	10.5 ; 14.4
	y2	20	11.31 (1.134)	11.51	9.5 ; 13.8
	y3	20	10.29 (1.214)	10.28	8.4 ; 12.8
Effect Size d*	y1	20	1.09 (0.092)	1.09	1.0 ; 1.3
	y2	20	1.09 (0.091)	1.12	0.9 ; 1.2
	y3	20	1.03 (0.103)	1.03	0.8 ; 1.2

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

Table 8 Continued.

Bayesian with Strata					
Sigma	y1	20	11.41 (0.521)	11.42	10.3 ; 12.1
	y2	20	10.60 (0.573)	10.58	9.8 ; 11.8
	y3	20	10.25 (0.492)	10.19	9.5 ; 11.2
Delta	y1	20	14.45 (0.990)	14.30	13.0 ; 16.8
	y2	20	13.41 (1.180)	13.36	11.4 ; 15.9
	y3	20	12.31 (1.210)	12.10	10.6 ; 14.9
Effect Size d*	y1	20	1.27 (0.089)	1.27	1.1 ; 1.4
	y2	20	1.27 (0.082)	1.29	1.1 ; 1.4
	y3	20	1.21 (0.092)	1.20	1.1 ; 1.3
Bayesian with Strata and Power Prior 0.25					
Sigma	y1	20	11.20 (0.509)	11.22	10.1 ; 11.9
	y2	20	10.45 (0.574)	10.41	9.6 ; 11.6
	y3	20	10.09 (0.479)	9.99	9.4 ; 11.1
Delta	y1	20	12.81 (0.828)	12.84	11.5 ; 14.7
	y2	20	12.03 (1.259)	12.14	10.0 ; 14.7
	y3	20	11.08 (1.301)	10.92	9.0 ; 13.8
Effect Size d*	y1	20	1.15 (0.089)	1.14	1.0 ; 1.3
	y2	20	1.16 (0.096)	1.19	0.9 ; 1.3
	y3	20	1.10 (0.109)	1.10	0.9 ; 1.3
Bayesian with Strata and Power Prior 0.5					
Sigma	y1	20	11.10 (0.507)	11.09	10.1 ; 11.8
	y2	20	10.34 (0.559)	10.25	9.6 ; 11.5
	y3	20	10.02 (0.497)	9.98	9.3 ; 11.0
Delta	y1	20	13.27 (0.868)	13.24	11.8 ; 15.0
	y2	20	12.44 (1.300)	12.51	10.3 ; 15.3
	y3	20	11.46 (1.375)	11.22	9.2 ; 14.2
Effect Size d*	y1	20	1.20 (0.090)	1.20	1.1 ; 1.4
	y2	20	1.21 (0.103)	1.23	1.0 ; 1.4
	y3	20	1.15 (0.118)	1.14	0.9 ; 1.3

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

APPENDIX D

SIMULATION RESULTS, SCENARIO 3

Table 9. Scenario 3, Input Data Set Parameter Characteristics

		Statistics			
Parameter	Variable	N	Mean (SD)	Median	95% CI
Placebo Mean	y1	20	10.3 (0.74)	10.3	9.12; 11.81
	y2	20	10.4 (0.79)	10.5	8.54; 11.82
	y3	20	10.1 (0.74)	10.0	9.17; 11.84
Sigma (STD)	y1	20	10.1 (0.50)	10.0	9.06; 11.08
	y2	20	10.1 (0.57)	10.1	9.13; 10.90
	y3	20	10.0 (0.47)	9.9	9.17; 10.82
Delta	y1	20	10.5 (0.97)	10.8	8.39; 11.86
	y2	20	5.2 (0.84)	5.2	3.50; 6.69
	y3	20	0.1 (0.97)	0.3	-1.24; 1.43
Effect Size, d	y1	20	1.0 (0.10)	1.1	0.86; 1.21
	y2	20	0.5 (0.08)	0.5	0.36; 0.62
	y3	20	0.0 (0.10)	0.0	-0.12; 0.14

Table 10.Scenario 3, Effect Size Ordering of Input Data Sets

Order	Condition	Count	Percent (%)
1	d1>d2>d3	8000	100.0

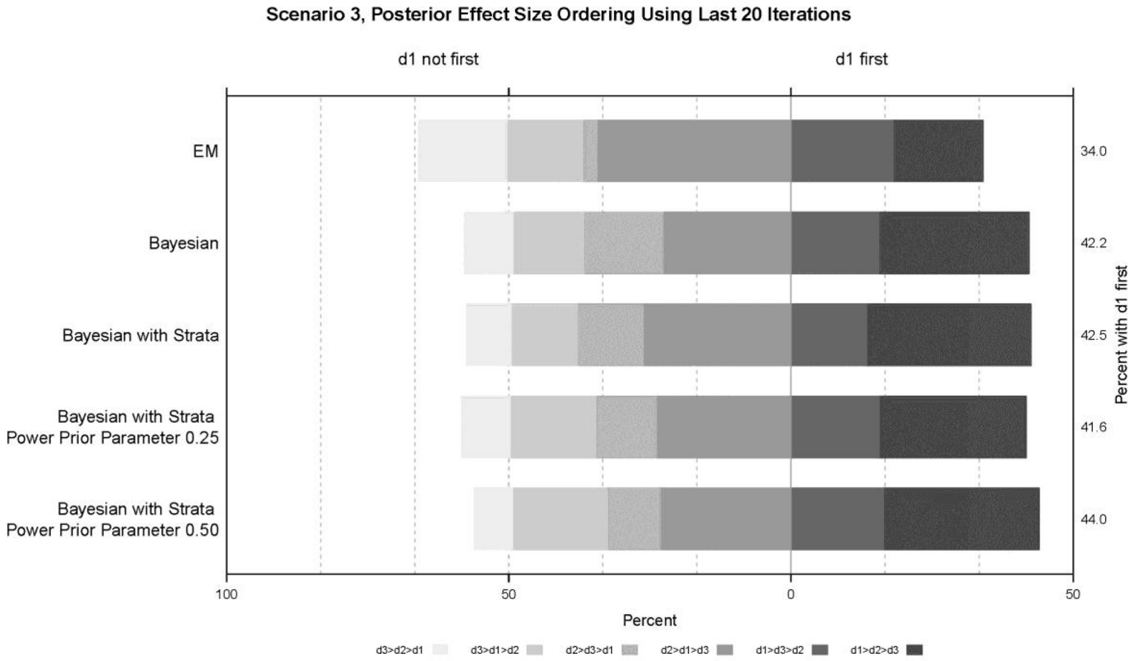


Figure 12. Scenario 3, Simulated Effect Size Ordering using Last 20 Iterations

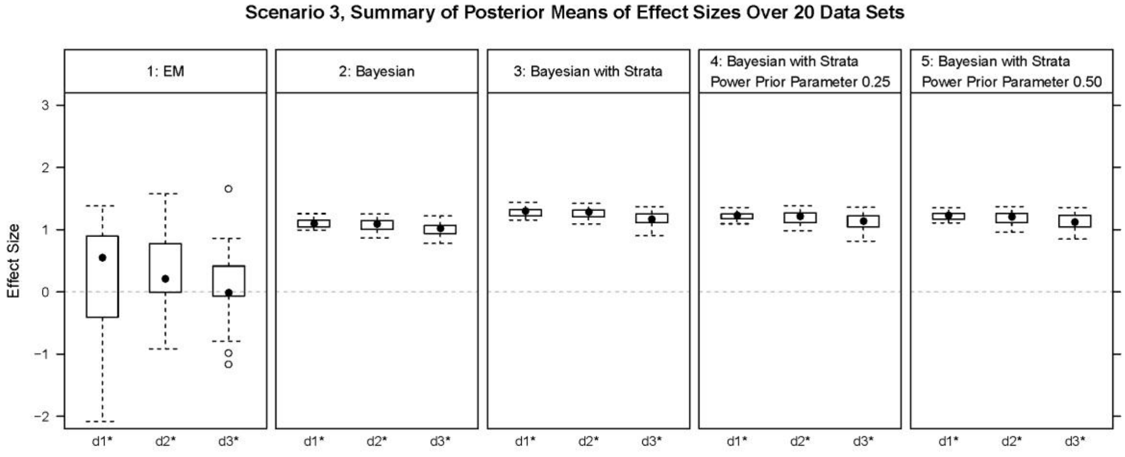


Figure 13. Scenario 3, Summary of Posterior Means of Effect Size Over 20 Data Sets

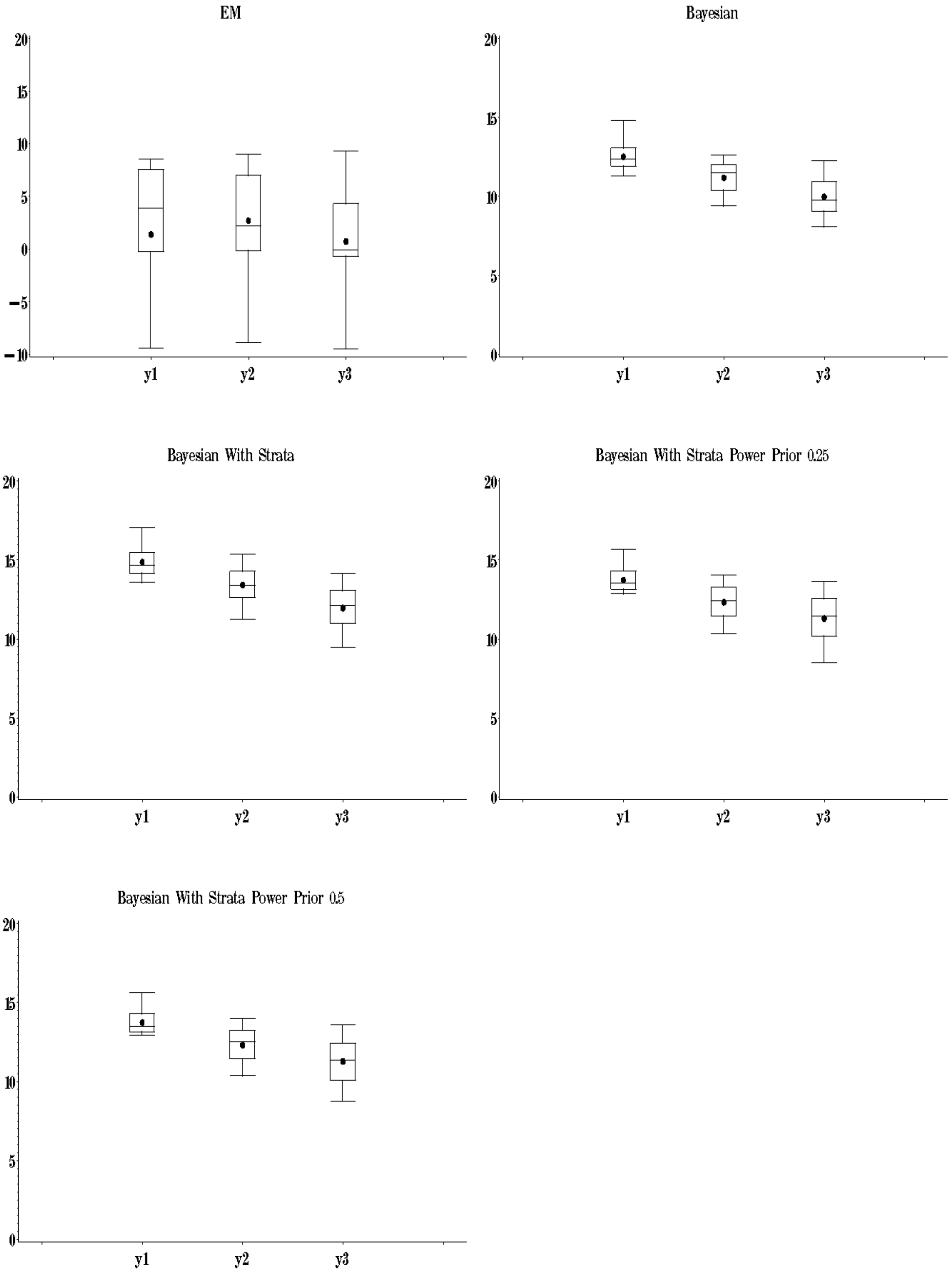


Figure 14. Scenario 3, Summary of Posterior Means of Delta Over 20 Data Sets

Table 11. Scenario 3, Summary of Posterior Means of Parameters Over 20 Data Sets

Analysis Type Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
EM					
Sigma	y1	19	8.91 (2.763)	9.27	0.2 ; 11.9
	y2	16	9.29 (1.650)	9.59	5.7 ; 11.2
	y3	17	9.04 (1.003)	9.29	7.4 ; 10.3
Delta	y1	19	1.39 (6.891)	2.34	-11.1 ; 8.5
	y2	16	2.68 (4.948)	2.20	-8.9 ; 9.0
	y3	17	0.72 (5.280)	-0.12	-9.5 ; 9.3
Effect Size d*	y1	19	1.41 (5.760)	0.55	-2.1 ; 24.9
	y2	16	0.38 (0.659)	0.21	-0.9 ; 1.6
	y3	17	0.11 (0.705)	-0.01	-1.2 ; 1.7
Bayesian					
Sigma	y1	20	11.42 (0.531)	11.53	10.6 ; 12.2
	y2	20	10.44 (0.670)	10.33	9.5 ; 11.5
	y3	20	10.02 (0.478)	10.02	9.2 ; 10.9
Delta	y1	20	12.51 (0.794)	12.32	11.3 ; 14.8
	y2	20	11.18 (1.037)	11.50	9.4 ; 12.6
	y3	20	9.98 (1.220)	9.76	8.1 ; 12.2
Effect Size d*	y1	20	1.10 (0.074)	1.11	1.0 ; 1.3
	y2	20	1.08 (0.100)	1.10	0.9 ; 1.3
	y3	20	1.00 (0.112)	1.03	0.8 ; 1.2

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

Table 11 Continued.

Bayesian with Strata					
Sigma	y1	20	11.63 (0.579)	11.71	10.7 ; 12.4
	y2	20	10.66 (0.661)	10.65	9.7 ; 11.9
	y3	20	10.19 (0.499)	10.23	9.3 ; 11.2
Delta	y1	20	14.87 (0.880)	14.66	13.6 ; 17.0
	y2	20	13.42 (1.077)	13.39	11.2 ; 15.4
	y3	20	11.96 (1.321)	12.10	9.5 ; 14.1
Effect Size d*	y1	20	1.29 (0.076)	1.30	1.2 ; 1.4
	y2	20	1.27 (0.082)	1.29	1.1 ; 1.4
	y3	20	1.18 (0.109)	1.17	0.9 ; 1.4
Bayesian with Strata and Power Prior 0.25					
Sigma	y1	20	11.32 (0.556)	11.39	10.3 ; 12.1
	y2	20	10.39 (0.674)	10.32	9.4 ; 11.6
	y3	20	10.01 (0.489)	10.04	9.2 ; 10.9
Delta	y1	20	13.73 (0.749)	13.54	12.8 ; 15.6
	y2	20	12.33 (1.181)	12.40	10.3 ; 14.0
	y3	20	11.30 (1.460)	11.43	8.5 ; 13.6
Effect Size d*	y1	20	1.22 (0.072)	1.24	1.1 ; 1.4
	y2	20	1.19 (0.104)	1.22	1.0 ; 1.4
	y3	20	1.14 (0.135)	1.14	0.8 ; 1.4
Bayesian with Strata and Power Prior 0.5					
Sigma	y1	20	11.33 (0.530)	11.43	10.5 ; 12.1
	y2	20	10.41 (0.683)	10.35	9.5 ; 11.6
	y3	20	10.01 (0.492)	10.03	9.1 ; 10.9
Delta	y1	20	13.74 (0.743)	13.49	12.9 ; 15.6
	y2	20	12.30 (1.156)	12.52	10.3 ; 14.0
	y3	20	11.27 (1.465)	11.33	8.7 ; 13.6
Effect Size d*	y1	20	1.22 (0.070)	1.23	1.1 ; 1.4
	y2	20	1.19 (0.104)	1.21	1.0 ; 1.4
	y3	20	1.13 (0.134)	1.12	0.8 ; 1.4

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

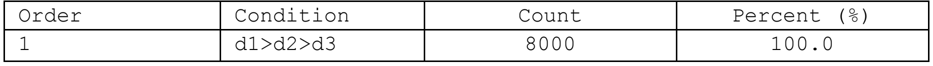
APPENDIX E

SIMULATION RESULTS, SCENARIO 4

Table 12. Scenario 4, Input Data Set Parameter Characteristics

Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
Placebo Mean	y1	20	10.7 (1.73)	10.7	7.19; 14.03
	y2	20	10.4 (1.42)	10.9	8.12; 13.07
	y3	20	10.0 (1.21)	9.7	8.13; 12.73
Sigma (STD)	y1	20	9.9 (0.95)	10.1	8.50; 11.64
	y2	20	10.0 (0.84)	10.0	8.62; 11.66
	y3	20	9.9 (0.96)	9.9	8.38; 11.56
Delta	y1	20	11.3 (1.92)	11.1	8.12; 14.1
	y2	20	5.0 (1.09)	4.8	3.48; 7.28
	y3	20	0.0 (0.60)	0.1	-1.02; 0.87
Effect Size, d	y1	20	1.1 (0.20)	1.1	0.89;1.6
	y2	20	0.5 (0.09)	0.5	0.36; 0.64
	y3	20	0.00 (0.06)	0.0	-0.11; 0.09

Table 13. Scenario 4, Effect Size Ordering of Input Data Sets



Order	Condition	Count	Percent (%)
1	d1>d2>d3	8000	100.0

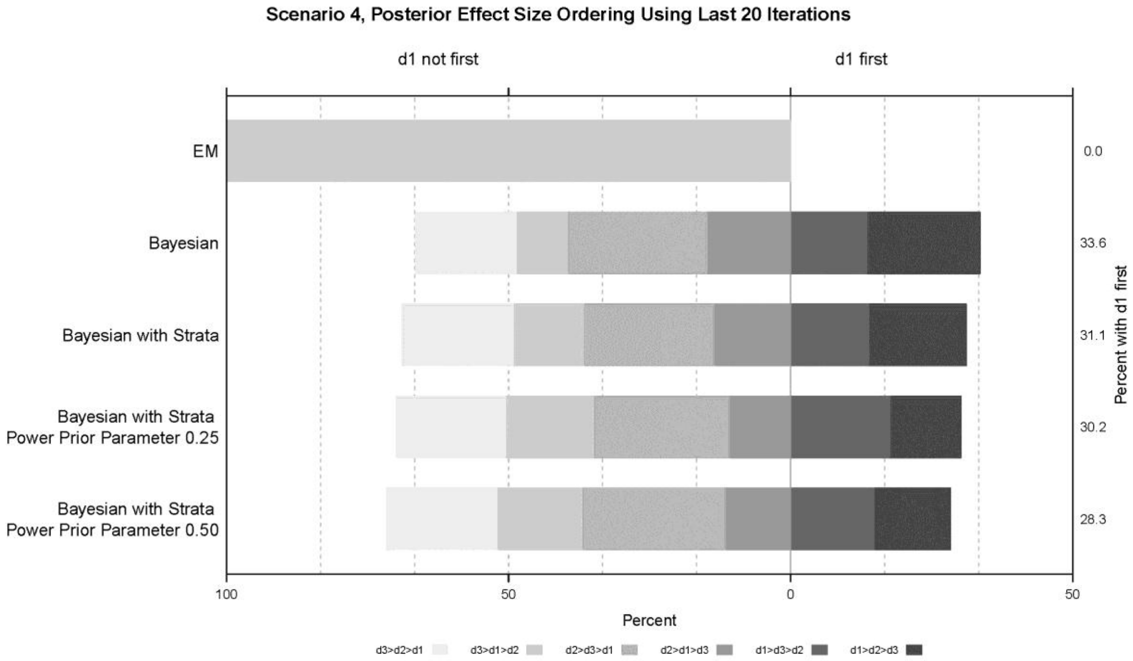


Figure 15. Scenario 4, Simulated Effect Size Ordering using Last 20 Iterations

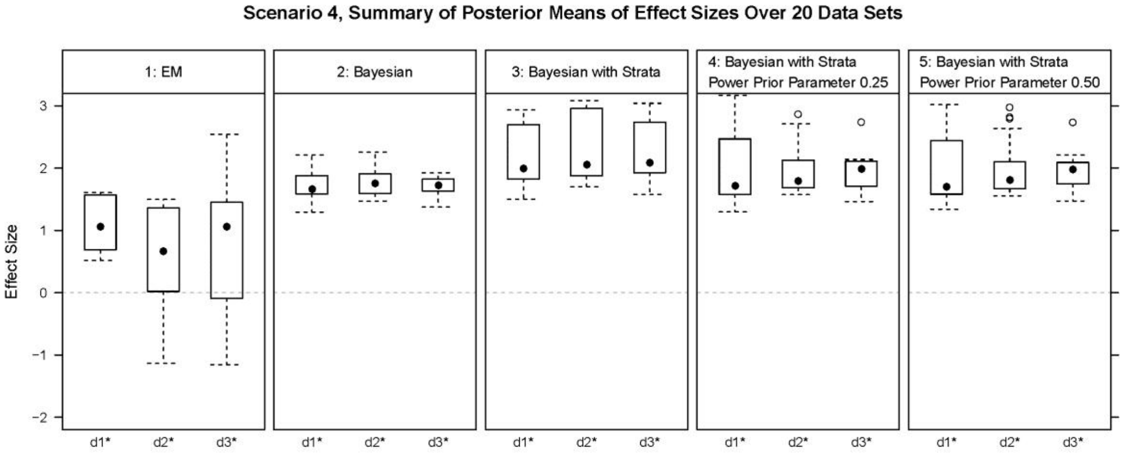


Figure 16. Scenario 4, Summary of Posterior Means of Effect Size Over 20 Data Sets

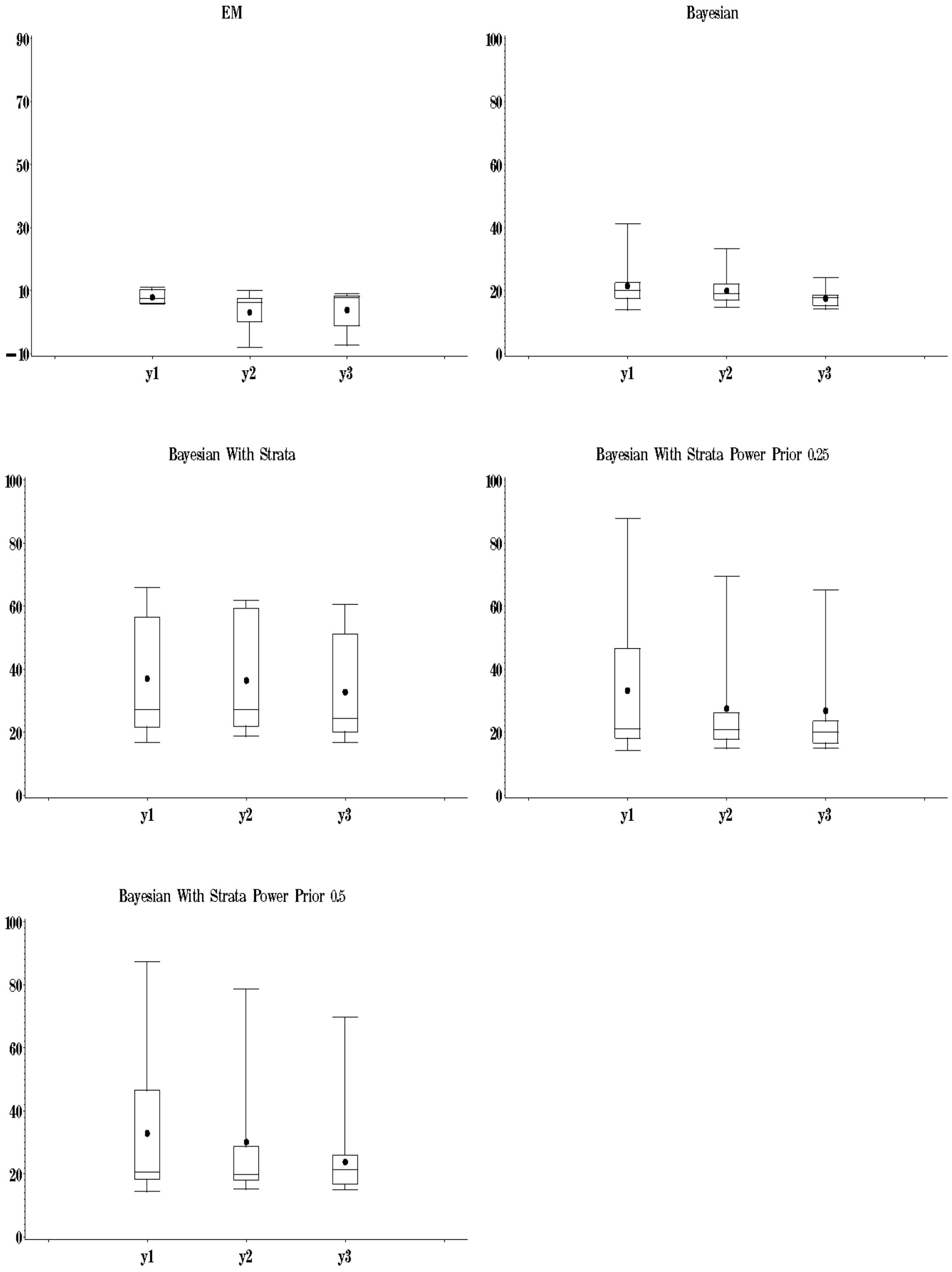


Figure 17. Scenario 4, Summary of Posterior Means of Delta Over 20 Data Sets

Table 14. Scenario 4, Summary of Posterior Means of Parameters Over 20 Data Sets

Analysis Type Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
EM					
Sigma	y1	6	8.02 (2.305)	8.04	4.7 ; 11.8
	y2	5	7.87 (2.224)	7.33	4.9 ; 10.8
	y3	8	6.62 (1.170)	6.50	5.1 ; 8.2
Delta	y1	6	7.87 (2.137)	7.31	5.6 ; 10.8
	y2	5	3.12 (7.209)	6.12	-8.1 ; 10.0
	y3	8	3.86 (7.058)	7.66	-7.5 ; 8.8
Effect Size d*	y1	6	1.08 (0.466)	1.06	0.5 ; 1.6
	y2	5	0.48 (1.081)	0.67	-1.1 ; 1.5
	y3	8	0.78 (1.292)	1.06	-1.2 ; 2.5
Bayesian					
Sigma	y1	20	12.44 (1.985)	12.08	10.2 ; 18.8
	y2	20	11.26 (1.302)	10.99	9.5 ; 14.8
	y3	20	10.31 (1.149)	10.23	8.2 ; 12.6
Delta	y1	20	21.49 (6.312)	19.94	13.8 ; 41.1
	y2	20	19.95 (4.254)	18.97	14.7 ; 33.1
	y3	20	17.49 (2.519)	17.71	14.0 ; 23.9
Effect Size d*	y1	20	1.72 (0.239)	1.67	1.3 ; 2.2
	y2	20	1.78 (0.213)	1.76	1.5 ; 2.3
	y3	20	1.72 (0.149)	1.73	1.4 ; 1.9

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

Table 14 Continued.

Bayesian with Strata					
Sigma	y1	20	16.06 (4.498)	14.07	11.1 ; 24.0
	y2	20	15.19 (3.860)	13.89	11.0 ; 21.0
	y3	20	13.94 (3.939)	11.97	9.0 ; 21.3
Delta	y1	20	37.00 (18.54)	27.05	16.8 ; 65.7
	y2	20	36.45 (17.54)	27.02	18.7 ; 61.8
	y3	20	32.79 (16.53)	24.33	16.7 ; 60.4
Effect Size d*	y1	20	2.19 (0.502)	2.00	1.5 ; 2.9
	y2	20	2.30 (0.519)	2.06	1.7 ; 3.1
	y3	20	2.25 (0.476)	2.09	1.6 ; 3.0
Bayesian with Strata and Power Prior 0.25					
Sigma	y1	20	15.34 (6.310)	12.17	10.4 ; 29.7
	y2	20	13.40 (4.908)	11.24	9.4 ; 25.8
	y3	20	12.56 (5.190)	10.94	8.4 ; 29.9
Delta	y1	20	33.27 (24.89)	21.02	14.1 ; 87.6
	y2	20	27.55 (17.12)	20.60	14.8 ; 69.3
	y3	20	26.81 (20.66)	20.56	14.8 ; 101.6
Effect Size d*	y1	20	1.97 (0.594)	1.72	1.3 ; 3.2
	y2	20	1.95 (0.385)	1.79	1.6 ; 2.9
	y3	20	2.00 (0.430)	1.99	1.5 ; 3.4
Bayesian with Strata and Power Prior 0.5					
Sigma	y1	20	15.28 (6.230)	12.20	10.3 ; 29.7
	y2	20	14.10 (5.532)	11.76	9.5 ; 26.3
	y3	20	11.92 (3.635)	10.89	8.7 ; 25.8
Delta	y1	20	32.88 (24.21)	20.46	14.3 ; 87.2
	y2	20	30.15 (20.32)	19.83	15.0 ; 78.6
	y3	20	23.81 (11.83)	21.32	14.9 ; 69.5
Effect Size d*	y1	20	1.96 (0.574)	1.70	1.3 ; 3.0
	y2	20	1.99 (0.451)	1.81	1.6 ; 3.0
	y3	20	1.95 (0.273)	1.99	1.5 ; 2.7

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

APPENDIX F

SIMULATION RESULTS, SCENARIO 5

Table 15. Scenario 5, Input Data Set Parameter Characteristics

		Statistics			
Parameter	Variable	N	Mean (SD)	Median	95% CI
Placebo Mean	y1	20	10.1 (0.44)	10.1	8.97; 10.96
	y2	20	10.0 (0.49)	10.1	9.13; 10.87
	y3	20	10.1 (0.30)	10.0	9.61; 10.57
Sigma (STD)	y1	20	10.1 (0.31)	10.1	9.54; 10.76
	y2	20	10.0 (0.36)	10.0	9.29; 10.70
	y3	20	9.9 (0.40)	9.8	9.11; 10.70
Delta	y1	20	10.1 (0.78)	10.0	9.05; 11.41
	y2	20	4.8 (0.73)	4.9	3.79; 6.49
	y3	20	0.1 (0.71)	-0.3	-1.23; 1.42
Effect Size, d	y1	20	1.0 (0.08)	1.0	0.86; 1.17
	y2	20	0.5 (0.08)	0.5	0.35; 0.64
	y3	20	-0.0 (0.07)	-0.0	-0.13; 0.14

Table 16.Scenario 5, Effect Size Ordering of Input Data Sets

Order	Condition	Count	Percent (%)
1	d1>d2>d3	8000	100.0

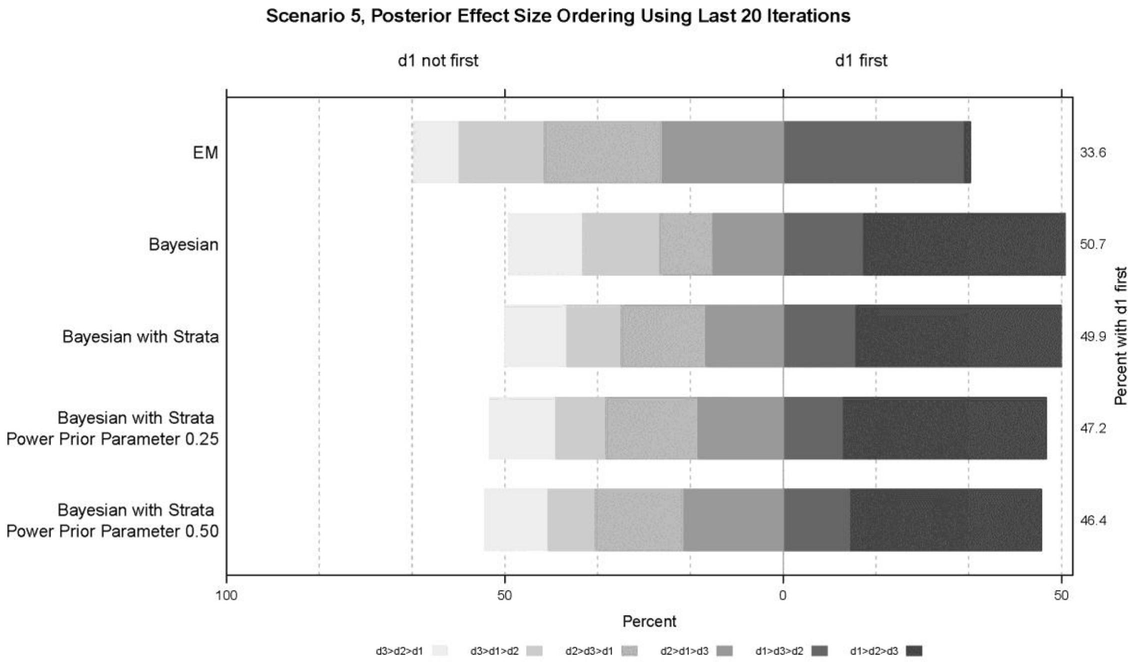


Figure 18. Scenario 5, Simulated Effect Size Ordering using Last 20 Iterations

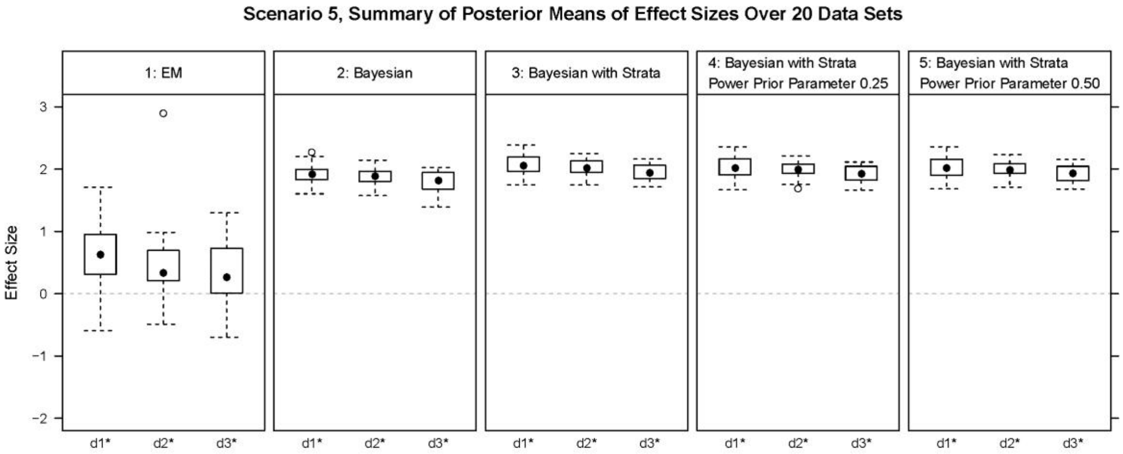


Figure 19. Scenario 5, Summary of Posterior Means of Effect Size Over 20 Data Sets

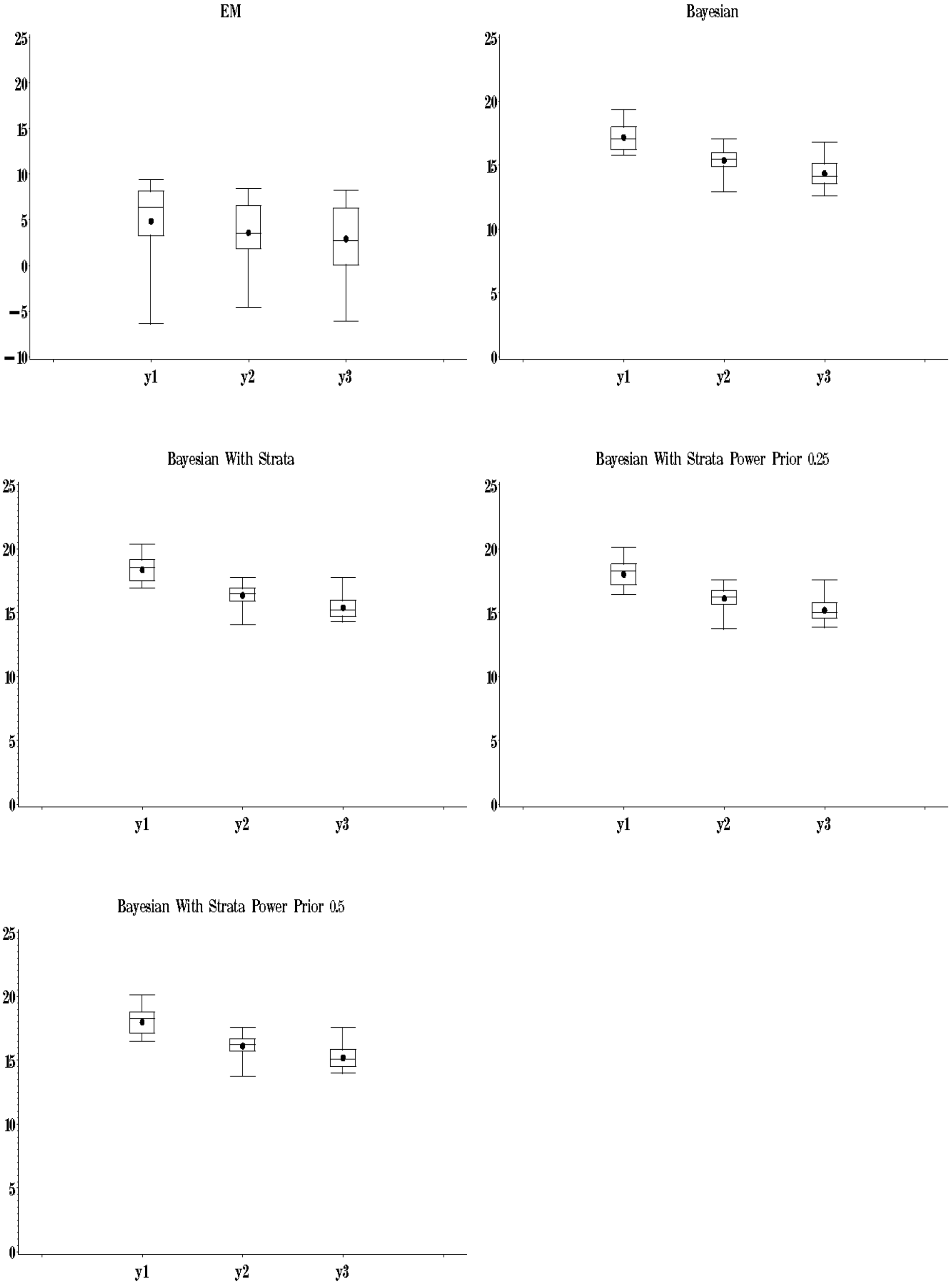


Figure 20. Scenario 5, Summary of Posterior Means of Delta Over 20 Data Sets

Table 17. Scenario 5, Summary of Posterior Means of Parameters Over 20 Data Sets

Analysis Type Parameter	Variable	Statistics			
		N	Mean (SD)	Median	95% CI
EM					
Sigma	y1	16	9.54 (1.688)	10.01	5.2 ; 11.4
	y2	15	9.19 (1.580)	9.47	4.0 ; 10.5
	y3	18	8.99 (1.163)	9.58	6.1 ; 10.1
Delta	y1	16	4.85 (4.493)	6.36	-6.4 ; 9.3
	y2	15	3.61 (3.460)	3.52	-4.6 ; 8.4
	y3	18	2.90 (3.969)	2.65	-6.1 ; 8.2
Effect Size d*	y1	16	0.60 (0.580)	0.63	-0.6 ; 1.7
	y2	15	0.54 (0.755)	0.34	-0.5 ; 2.9
	y3	18	0.37 (0.527)	0.27	-0.7 ; 1.3
Bayesian					
Sigma	y1	20	8.96 (0.465)	8.90	8.1 ; 9.8
	y2	20	8.20 (0.340)	8.21	7.3 ; 8.8
	y3	20	8.03 (0.446)	7.92	7.1 ; 9.1
Delta	y1	20	17.16 (1.003)	17.03	15.8 ; 19.3
	y2	20	15.36 (0.897)	15.44	12.9 ; 17.0
	y3	20	14.33 (1.124)	14.11	12.6 ; 16.7
Effect Size d*	y1	20	1.93 (0.175)	1.92	1.6 ; 2.3
	y2	20	1.88 (0.146)	1.89	1.6 ; 2.1
	y3	20	1.80 (0.185)	1.82	1.4 ; 2.0

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

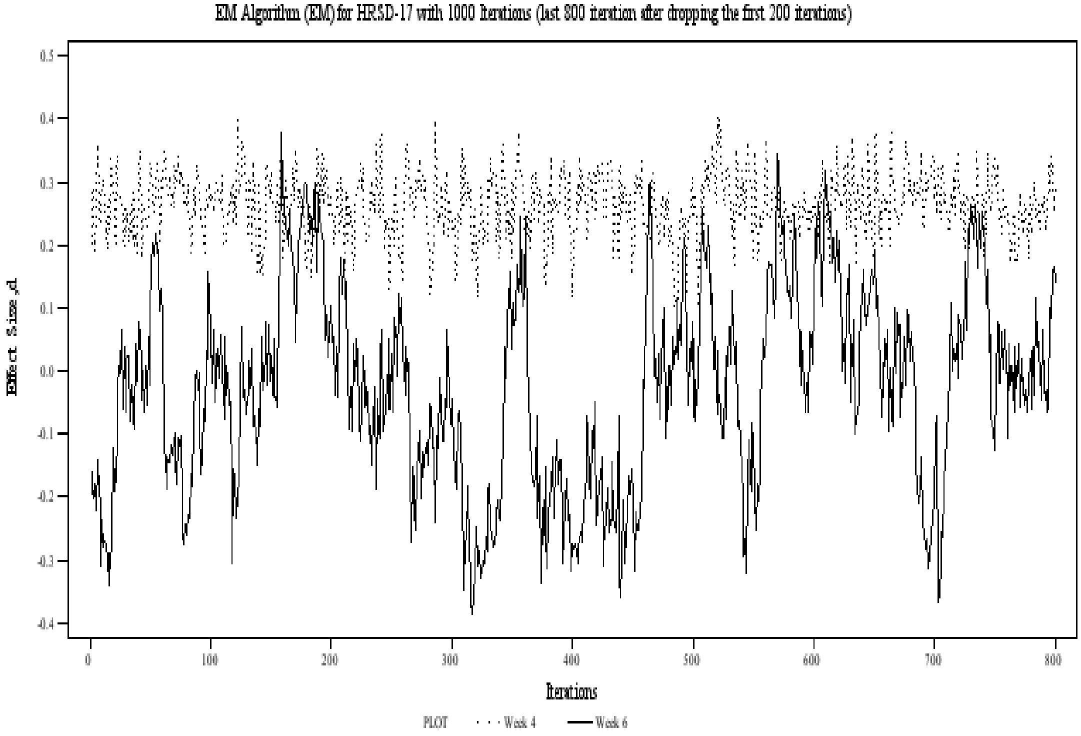
Table 17 Continued.

Bayesian with Strata					
Sigma	y1	20	8.89 (0.443)	8.69	8.1 ; 9.7
	y2	20	8.11 (0.307)	8.16	7.3 ; 8.7
	y3	20	7.94 (0.386)	7.85	7.0 ; 8.7
Delta	y1	20	18.35 (0.976)	18.51	16.9 ; 20.3
	y2	20	16.36 (0.820)	16.44	14.0 ; 17.7
	y3	20	15.38 (0.907)	15.21	14.3 ; 17.7
Effect Size d*	y1	20	2.07 (0.163)	2.06	1.7 ; 2.4
	y2	20	2.02 (0.131)	2.02	1.7 ; 2.3
	y3	20	1.95 (0.134)	1.94	1.7 ; 2.2
Bayesian with Strata and Power Prior 0.25					
Sigma	y1	20	8.90 (0.466)	8.73	8.1 ; 9.8
	y2	20	8.13 (0.297)	8.14	7.3 ; 8.7
	y3	20	7.96 (0.397)	7.89	7.0 ; 8.6
Delta	y1	20	18.00 (1.038)	18.24	16.4 ; 20.1
	y2	20	16.11 (0.855)	16.19	13.7 ; 17.5
	y3	20	15.19 (0.937)	14.98	13.8 ; 17.5
Effect Size d*	y1	20	2.03 (0.176)	2.03	1.7 ; 2.4
	y2	20	1.99 (0.131)	2.00	1.7 ; 2.2
	y3	20	1.92 (0.142)	1.93	1.7 ; 2.1
Bayesian with Strata and Power Prior 0.5					
Sigma	y1	20	8.92 (0.467)	8.75	8.1 ; 9.8
	y2	20	8.13 (0.301)	8.16	7.3 ; 8.7
	y3	20	7.95 (0.395)	7.89	7.0 ; 8.7
Delta	y1	20	17.99 (1.025)	18.23	16.5 ; 20.1
	y2	20	16.12 (0.851)	16.17	13.7 ; 17.6
	y3	20	15.20 (0.943)	15.03	13.9 ; 17.6
Effect Size d*	y1	20	2.03 (0.174)	2.02	1.7 ; 2.4
	y2	20	1.99 (0.131)	1.99	1.7 ; 2.2
	y3	20	1.92 (0.145)	1.94	1.7 ; 2.2

CI= Credible Interval; Sigma= Pooled variance; Delta= Difference between treatment arms.

APPENDIX G

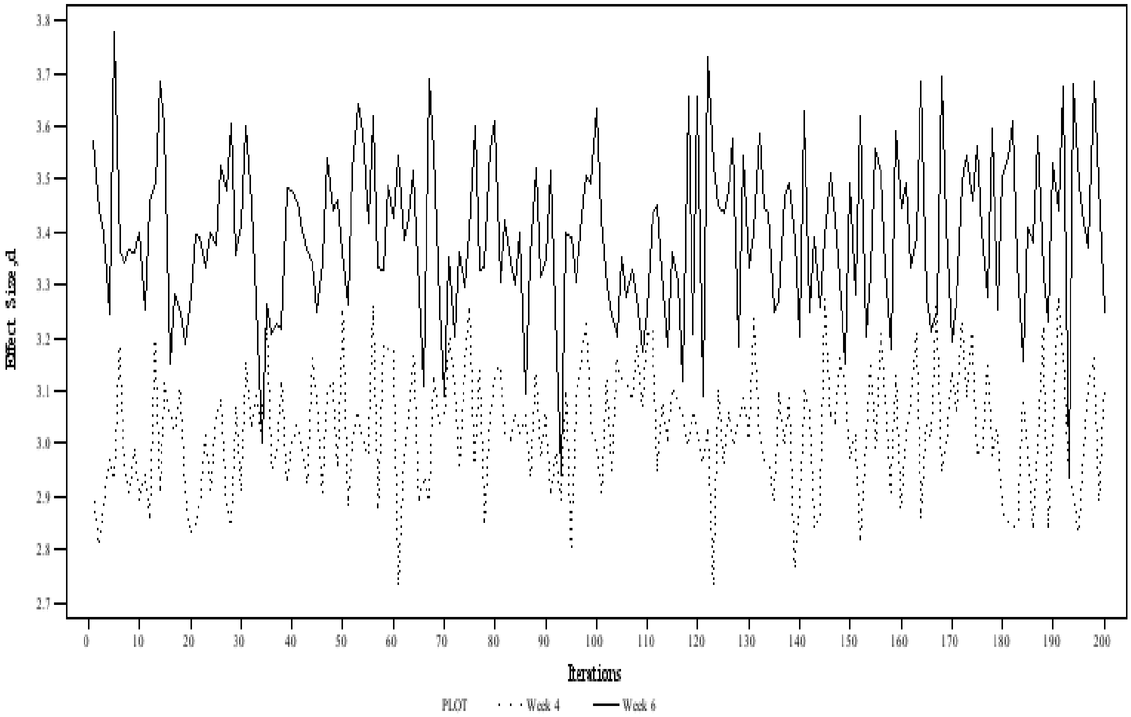
MDD STUDY RESULTS



Summary Statistics Week 4: n=800, mean=0.26, STD=0.050, range=(0.1, 0.4), median=0.27, 95%CI=(0.159,0.360)
Summary Statistics Week 6: n=800, mean=-0.03, STD=0.158, range=(-0.4, 0.4), median=-0.02, 95%CI=(-0.312,0.265)

Figure 21. EM – Posterior Inferences on Effect Size, HRSD-17

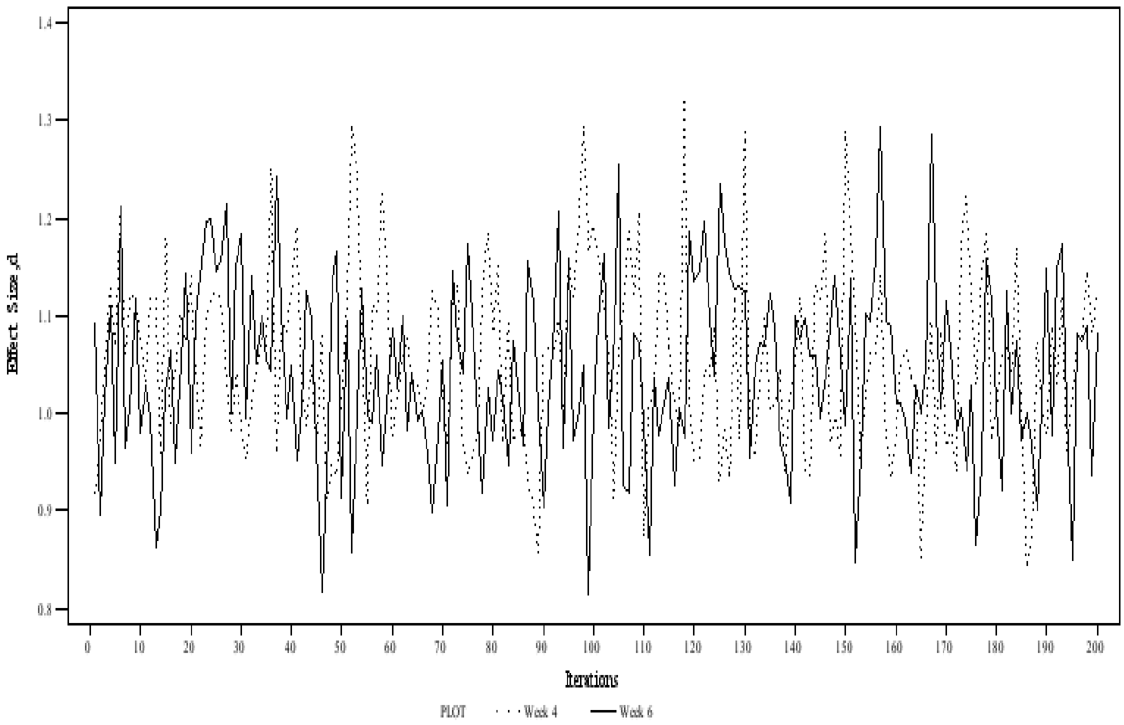
MCMC Algorithm (NOSTRATA) for HRSD-17 with 1000 Iterations (every 4th iteration after dropping the first 200 iterations)



Summary Statistics Week 4: n=200, mean=3.02, STD=0.119, range=(2.7, 3.3), median=3.02, 95%CI=(2.813,3.257)
Summary Statistics Week 6: n=200, mean=3.39, STD=0.154, range=(2.9, 3.8), median=3.39, 95%CI=(3.095,3.686)

Figure 22. Bayesian – Posterior Inferences on Effect Size, HRSD-17

MCMC Algorithm (strata) for HRSD-17 with 1000 Iterations (every 4th iteration after dropping the first 200 iterations)



Summary Statistics Week 4: n=200, mean=1.05, STD=0.092, range=(0.8, 1.3), median=1.05, 95%CI=(0.891,1.270)
Summary Statistics Week 6: n=200, mean=1.04, STD=0.092, range=(0.8, 1.3), median=1.04, 95%CI=(0.856,1.225)

Figure 23. Bayesian with Strata – Posterior Inferences on Effect Size, HRSD-17

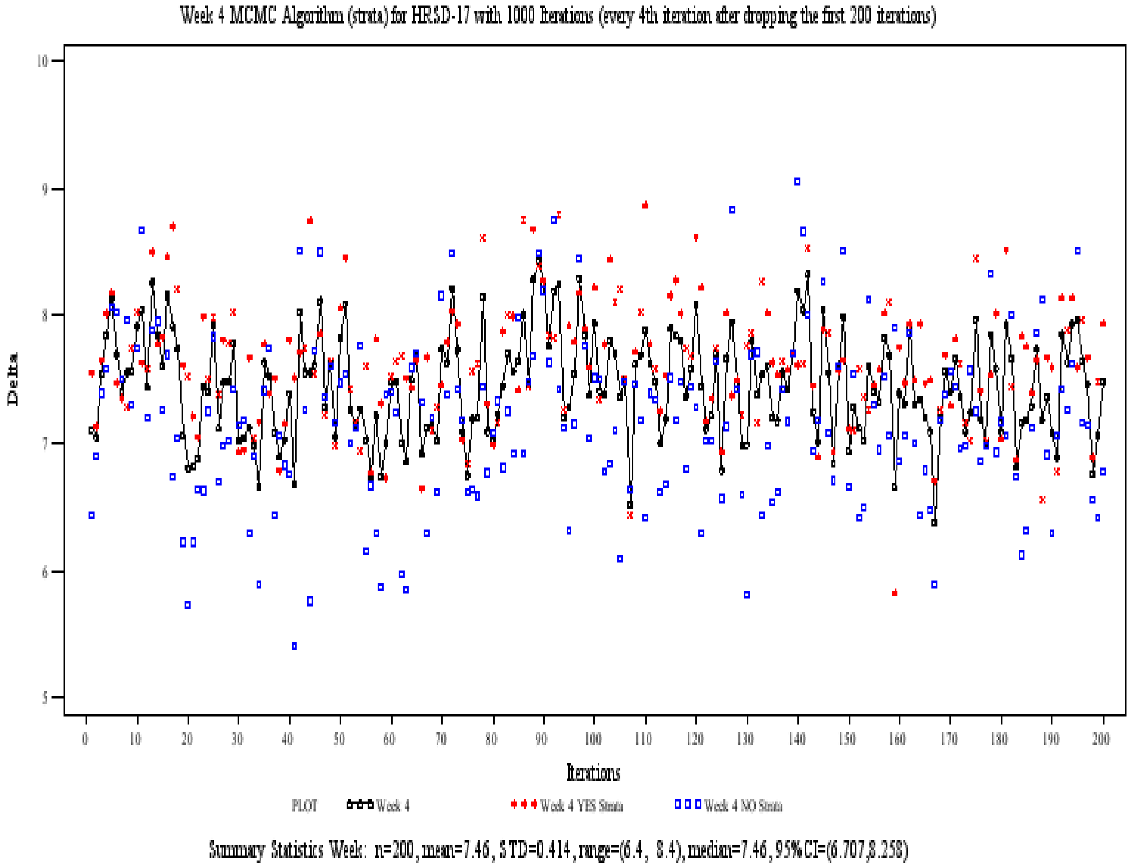


Figure 24. Bayesian with Strata at Week 4 – Posterior Inferences on Each Stratum on Mean Difference, HRSD-17

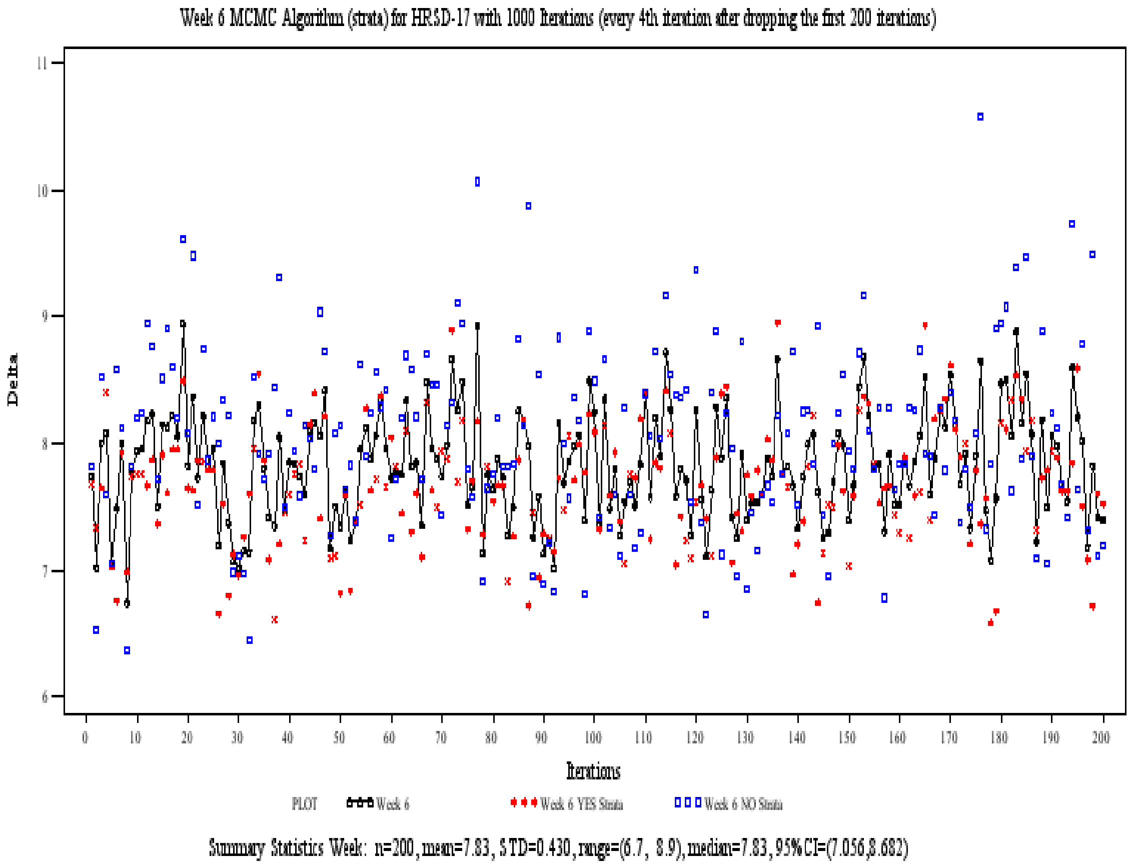
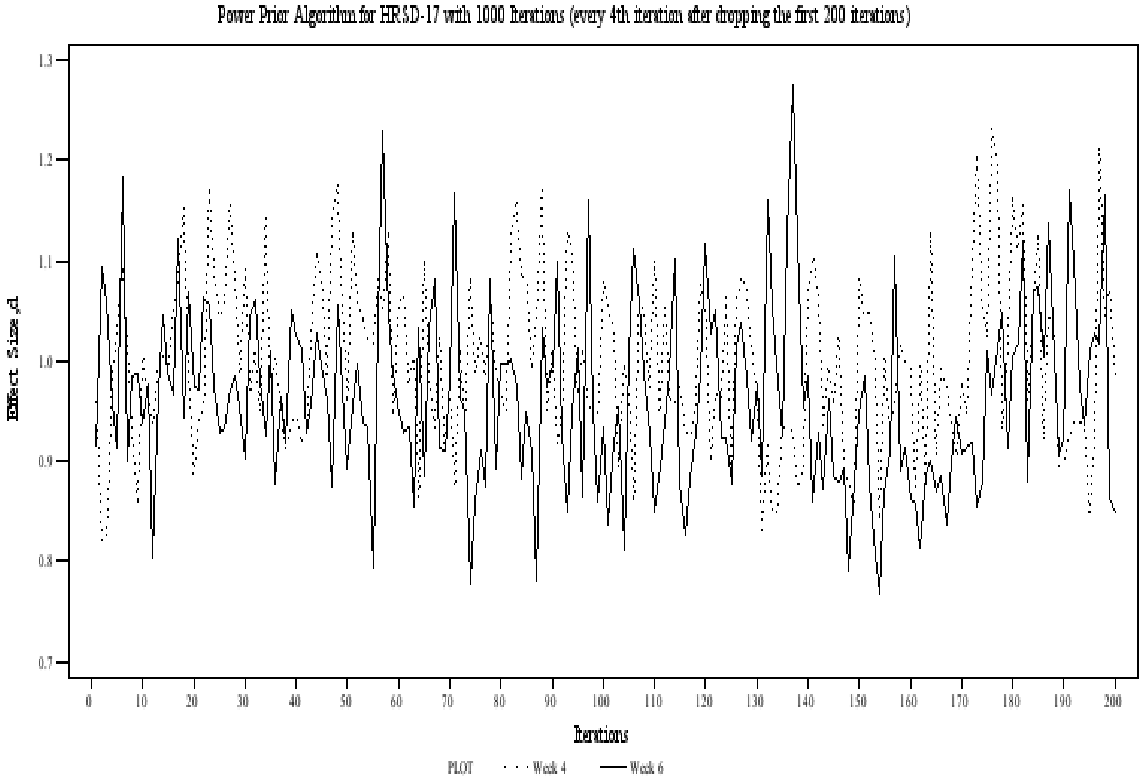


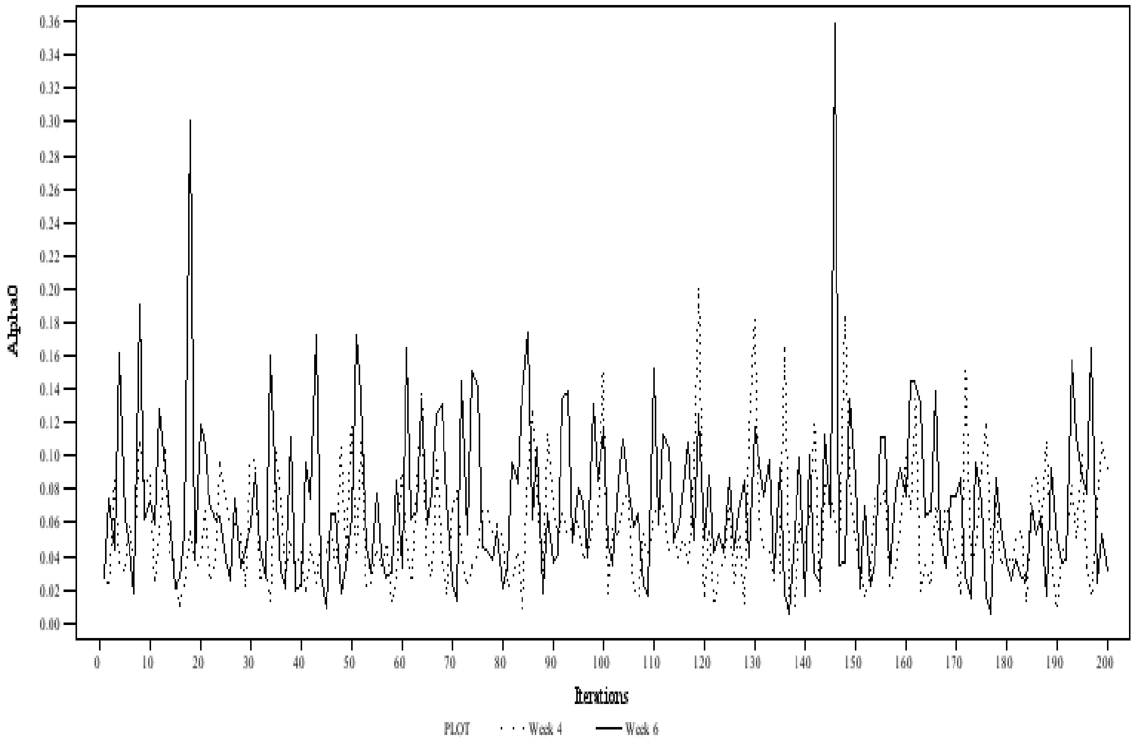
Figure 25. Bayesian with Strata at Week 6 – Posterior Inferences on Each Stratum on Mean Difference, HRSD-17



Summary Statistics Week 4: n=200, mean=1.00, STD=0.086, range=(0.8, 1.2), median=0.99, 95%CI=(0.846,1.176)
Summary Statistics Week 6: n=200, mean=0.96, STD=0.091, range=(0.8, 1.3), median=0.96, 95%CI=(0.798,1.166)

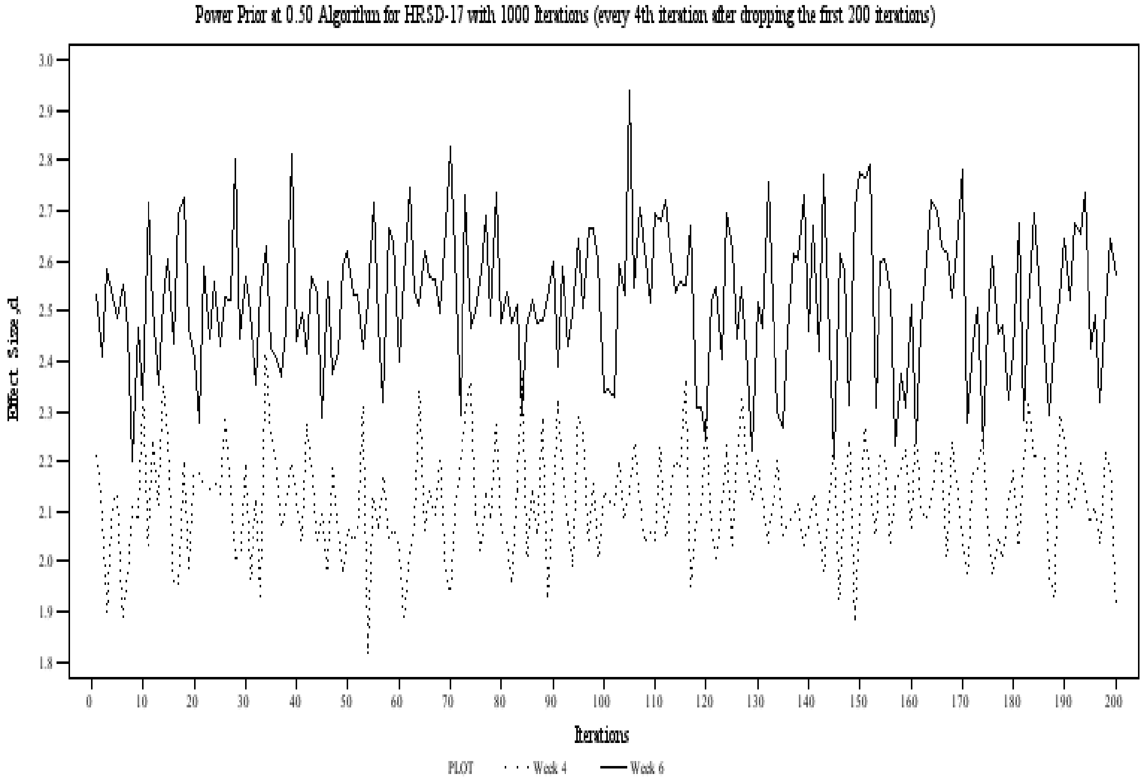
Figure 26. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on Effect Size, HRSD-17

Power Prior Algorithm for HRSD-17 with 1000 Iterations (every 4th iteration after dropping the first 200 iterations)



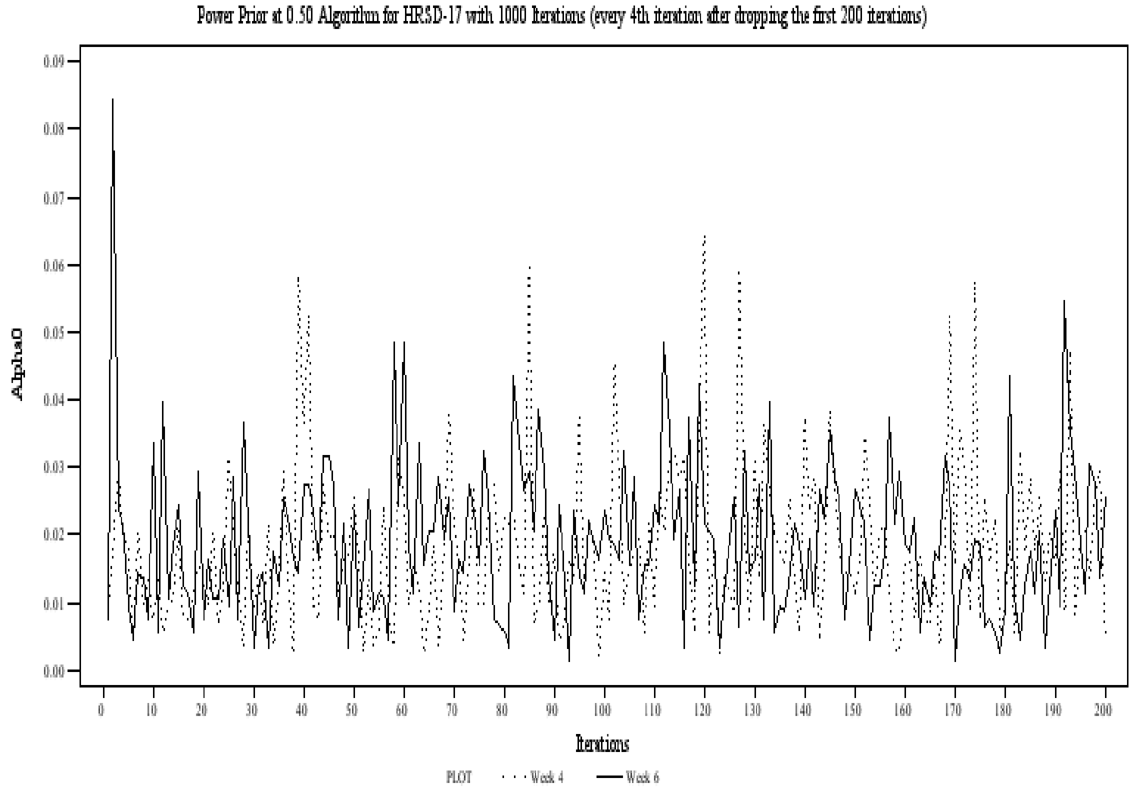
Summary Statistics Week 4: n=200, mean=0.06, STD=0.035, range=(0.0, 0.2), median=0.05, 95%CI=(0.010,0.151)
Summary Statistics Week 6: n=200, mean=0.07, STD=0.049, range=(0.0, 0.4), median=0.06, 95%CI=(0.015,0.172)

Figure 27. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on Power Parameter



Summary Statistics Week 4: n=200, mean=2.12, STD=0.111, range=(1.8, 2.4), median=2.12, 95%CI=(1.909,2.346)
Summary Statistics Week 6: n=200, mean=2.52, STD=0.143, range=(2.2, 2.9), median=2.52, 95%CI=(2.237,2.787)

Figure 28. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on Effect Size, HRSD-17

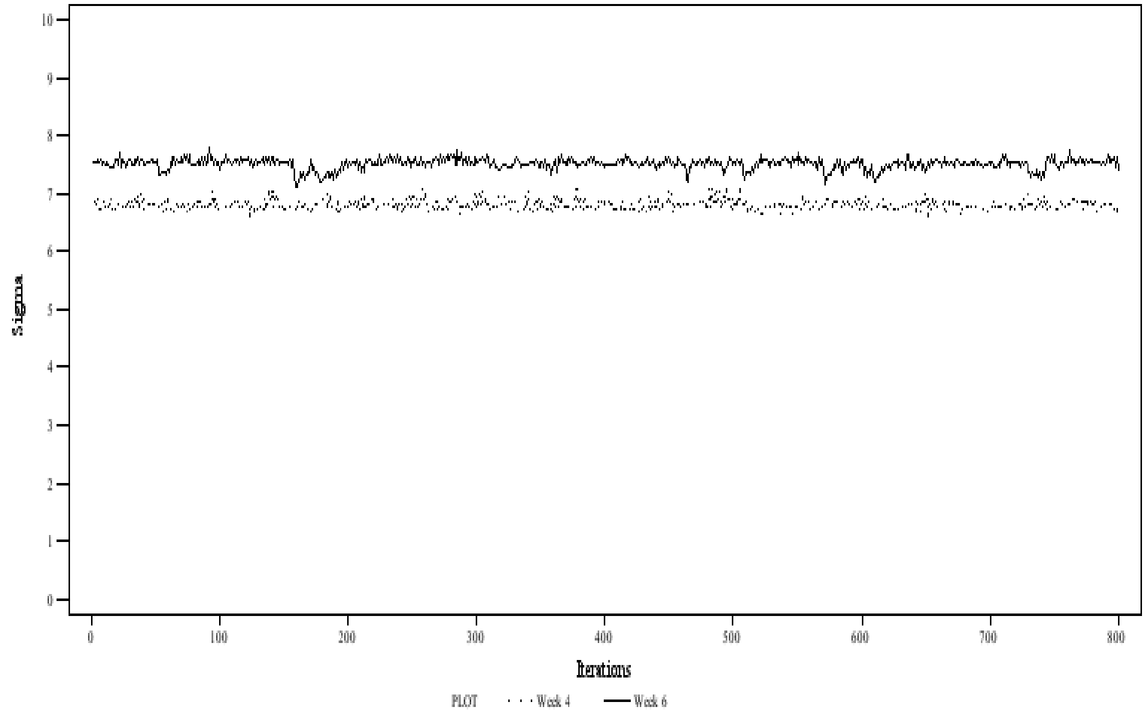


Power Prior at 0.50 Algorithm for HRSD-17 with 1000 Iterations (every 4th iteration after dropping the first 200 iterations)

Summary Statistics Week 4: n=200, mean=0.02, STD=0.012, range=(0.0, 0.1), median=0.02, 95%CI=(0.002,0.055)
Summary Statistics Week 6: n=200, mean=0.02, STD=0.012, range=(0.0, 0.1), median=0.02, 95%CI=(0.003,0.046)

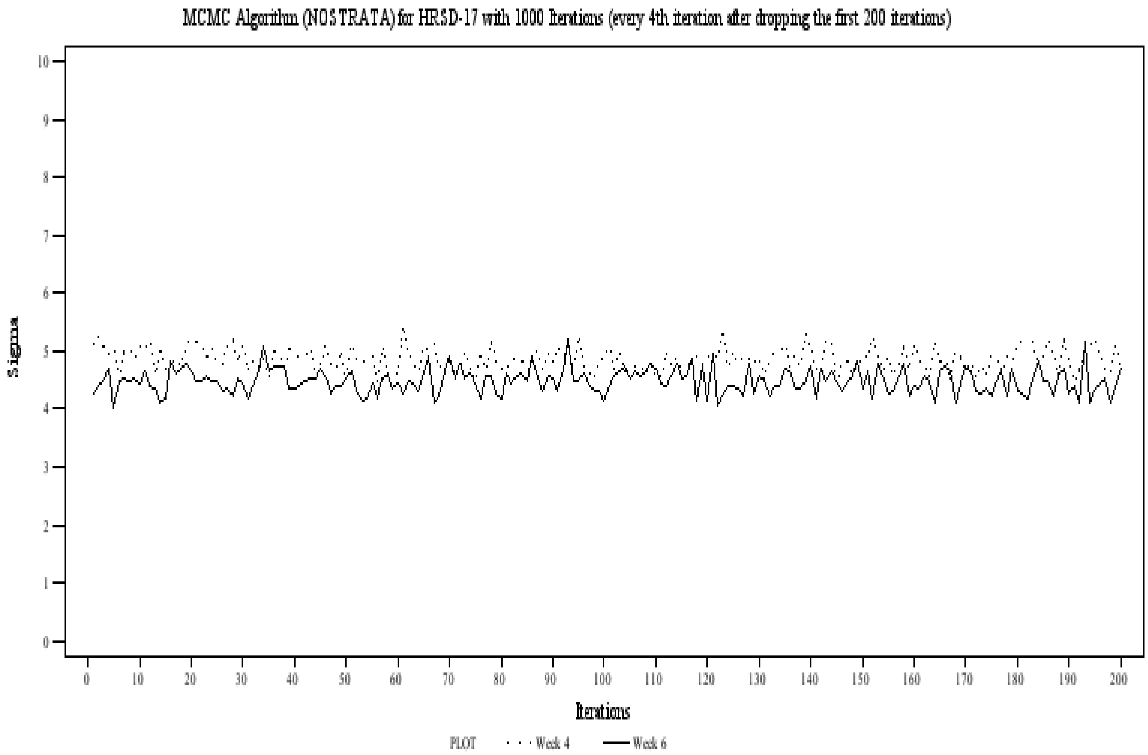
Figure 29. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on Power Parameter

EM Algorithm (EM) for HRSD-17 with 1000 Iterations (last 800 iteration after dropping the first 200 iterations)



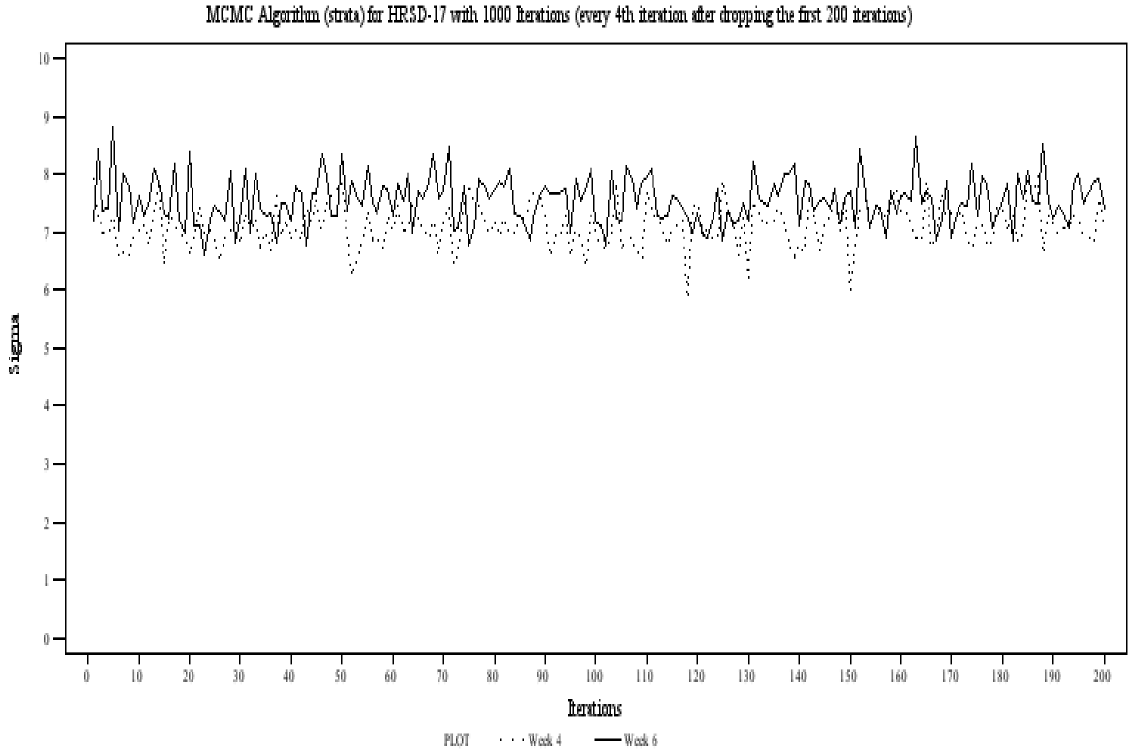
Summary Statistics Week 4: n=800, mean=6.81, STD=0.082, range=(6.6, 7.1), median=6.80, 95%CI=(6.686,7.011)
Summary Statistics Week 6: n=800, mean=7.52, STD=0.092, range=(7.1, 7.8), median=7.53, 95%CI=(7.279,7.670)

Figure 30. EM – Posterior Inferences on SD


MCMC Algorithm (NOSTRATA) for HRSD-17 with 1000 Iterations (every 4th iteration after dropping the first 200 iterations)

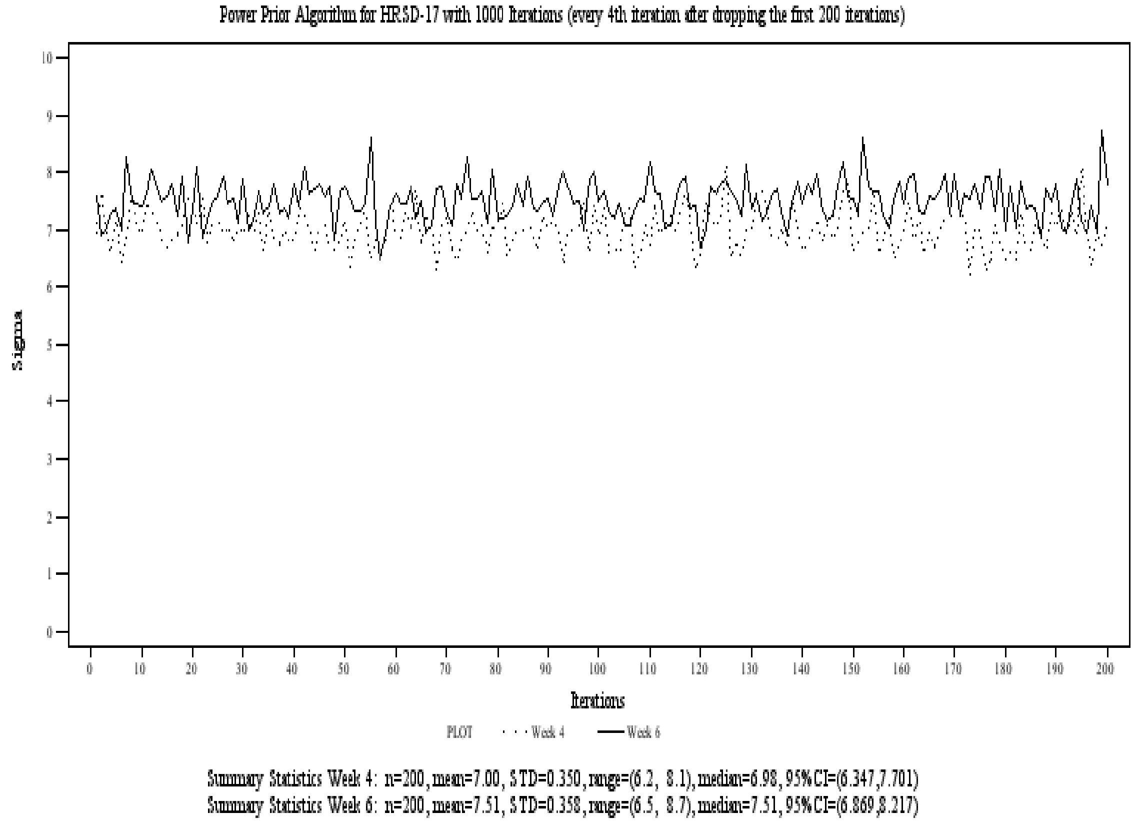
Summary Statistics Week 4: n=200, mean=4.87, STD=0.199, range=(4.5, 5.4), median=4.87, 95%CI=(4.513,5.255)
Summary Statistics Week 6: n=200, mean=4.49, STD=0.214, range=(4.0, 5.2), median=4.48, 95%CI=(4.102,4.928)

Figure 31. Bayesian – Posterior Inferences on SD



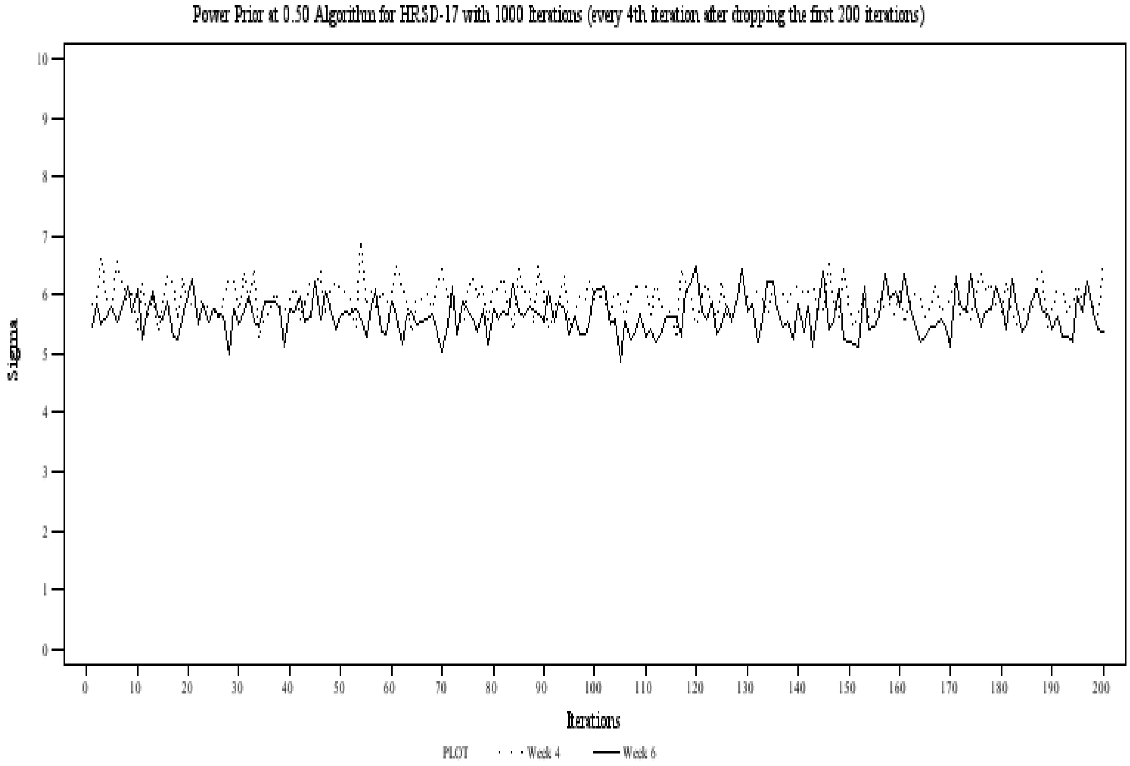
Summary Statistics Week 4: n=200, mean=7.12, STD=0.369, range=(5.9, 8.0), median=7.11, 95%CI=(6.442,7.858)
Summary Statistics Week 6: n=200, mean=7.55, STD=0.412, range=(6.6, 8.8), median=7.52, 95%CI=(6.815,8.423)

Figure 32. Bayesian with Strata – Posterior Inferences on SD



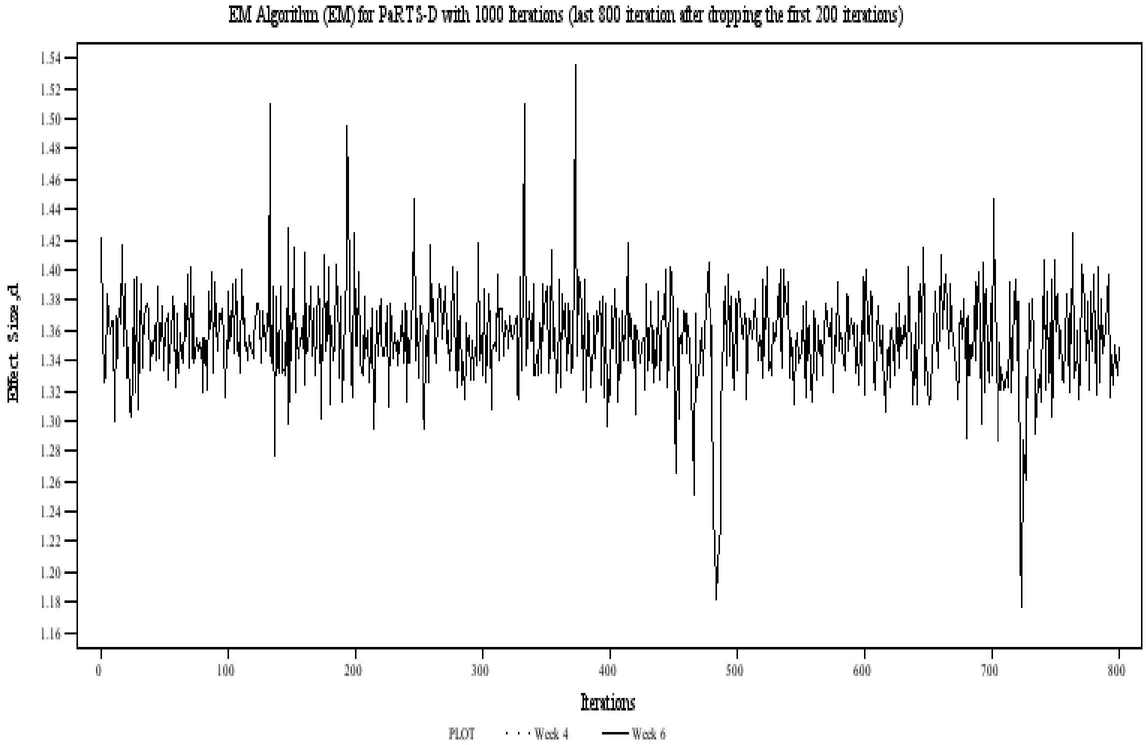
Summary Statistics Week 4: n=200, mean=7.00, STD=0.350, range=(6.2, 8.1), median=6.98, 95%CI=(6.347,7.701)
Summary Statistics Week 6: n=200, mean=7.51, STD=0.358, range=(6.5, 8.7), median=7.51, 95%CI=(6.869,8.217)

Figure 33. Bayesian with Strata and Power Prior 0.25 – Posterior Inferences on SD



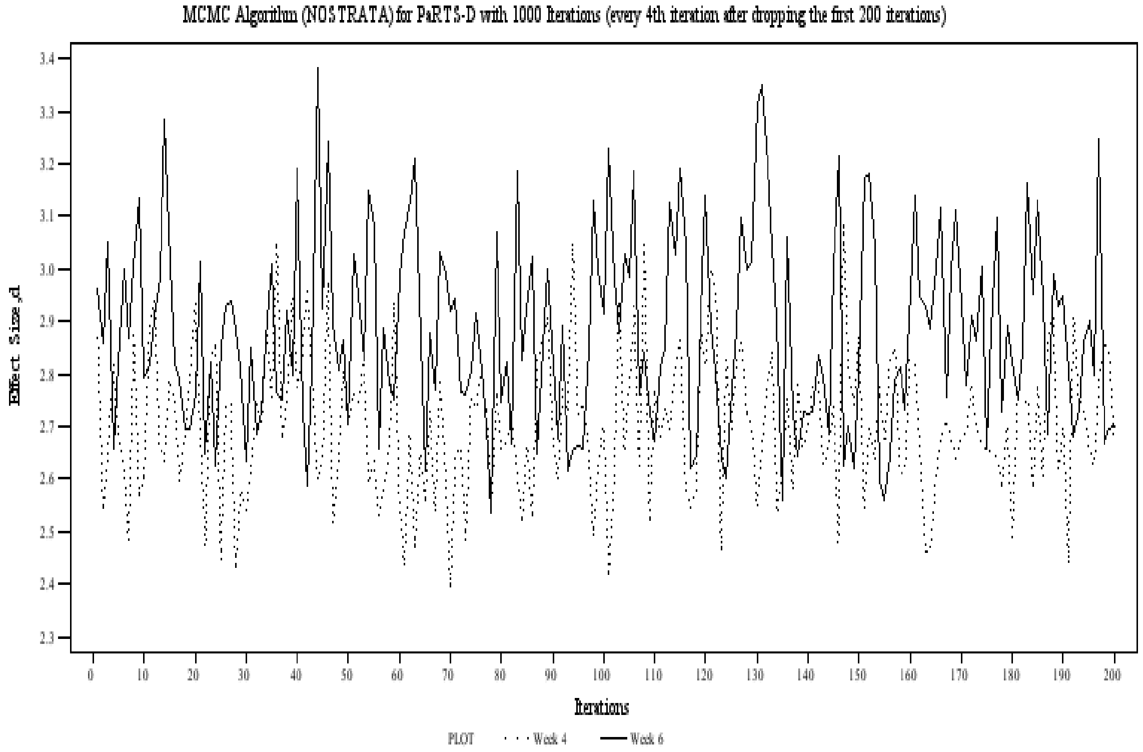
Summary Statistics Week 4: n=200, mean=5.94, STD=0.286, range=(5.3, 6.9), median=5.93, 95%CI=(5.418,6.500)
Summary Statistics Week 6: n=200, mean=5.67, STD=0.313, range=(4.9, 6.5), median=5.66, 95%CI=(5.127,6.363)

Figure 34. Bayesian with Strata and Power Prior 0.5 – Posterior Inferences on SD



Summary Statistics Week 4: n=0, mean=, STD=, range=(,), median=, 95%CI=(,)
Summary Statistics Week 6: n=800, mean=1.35, STD=0.033, range=(1.2, 1.5), median=1.35, 95%CI=(1.296,1.412)

Figure 35. PaRTS-D Effect Size Using EM Algorithm



Summary Statistics Week 4: n=200, mean=2.70, STD=0.138, range=(2.4, 3.1), median=2.70, 95% CI=(2.444,2.989)
Summary Statistics Week 6: n=200, mean=2.88, STD=0.179, range=(2.5, 3.4), median=2.86, 95% CI=(2.602,3.245)

Figure 36. PaRTS-D Effect Size Using Bayesian Approach