ABSTRACT

Bayesian Approaches for Design of Psychometric Studies
With Underreporting and Misclassification

Brandi Falley, Ph.D.

Chairperson: James D. Stamey, Ph.D.

Measurement error problems in binary regression are of considerable interest among researchers, especially in epidemiological studies. Misclassification can be considered a special case of measurement error specifically for the situation when measurement is the categorical classification of items. Bayesian methods offer practical advantages for the analysis of epidemiological data including the possibility of incorporating relevant prior scientific information and the ability to make inferences that do not rely on large sample assumptions.

Because of the high cost and time constraints for clinical trials, researchers often need to determine the smallest sample size that provides accurate inferences for a parameter of interest. Although most experimenters have employed frequentist methods, the Bayesian paradigm offers a wide variety of methodologies and are becoming increasingly more popular in clinical trials because of their flexibility and their ease of interpretation. We will simultaneously estimate efficacy and safety where the safety variable is subject to underreporting. We propose a Bayesian sample size determination method to account for the underreporting and appropriately power the study. We will allow efficacy and safety to be independent, as well as dependent using a regression model. For both models, we will allow the safety variable to be underreported.

Bayesian Approaches for Design of Psychometric Studies
With Underreporting and Misclassification

by

Brandi Falley, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

James D. Stamey, Ph.D, Chairperson

_____

A. Alexander Beajeaun, Ph.D.

_____

Thomas L. Bratcher, Ph.D.

_____

David Kahle, Ph.D.

_____

Dean M. Young, Ph.D.

Accepted by the Graduate School
December 2012

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to my advisor, Dr. James Stamey, for your constant support and guidance. Without you, I might still be trying to iron out my simulations. Thank you for always being available and easy to contact, even from a distance. Thank you to Dr. Jack Tubbs for encouraging me to attend Baylor. Thank you to Dr. Tom Bratcher for constantly pushing me to do better. To Dr. Alex Beaujean, thank you for all of your help on this dissertation, including the data set you supplied. To Dr. Jane Harvill, I can't thank you enough for your support and understanding of the things I went through during my time in graduate school. Thank you for allowing me to intrude unannounced and talk to you about what I was going through. I dealt with some life problems I didn't intend on dealing with, and your support definitely helped me keep pushing on and concentrating on school.

To my mentor at Baylor Health Care Systems in Dallas, Sunni Barnes, thank you. You were a fantastic leader who allowed me to constantly learn. To Lindsey Philpot, I greatly appreciate your friendship, your leadership, and your character as a colleague at BHCS, and especially on the Edgerton project. Thank you for welcoming me to the team with open arms. To the rest of the SRCT team at BHCS, thank you for accepting me and making me feel welcome.

To my mom and dad, thank you for your endless love and support up to and during my dissertation work. I love you both. Thank you to my colleagues, both past and present - Johnny, Monica, John, Bonnie, Stephanie, my old office-mate, Lindsay, and all the others who kept me laughing and sane these last four years.

Finally, thank you to Kevin for helping me through this adventure. Thank you for listening to all of my concerns and doubts, as well as my accomplishments and triumphs. Thank you for believing in me when I didn't believe in myself. Your love and support have pushed me through.

DEDICATION

To Kevin

My rock

Both near and far

My constant support

CHAPTER ONE

Introduction

In many research areas, and especially in social sciences, studies may involve variables that cannot be observed directly or are observed with error (Fox and Glas (2003)). Further, many forms of human response behavior are inherently stochastic in nature. Lord and Novick (1968) adhere to the so-called stochastic subject view in which it is assumed that responses of the subjects depend on small variations in the circumstances in which the response is generated. Accordingly, response variance is the variation in responses to the same question repeatedly administered to the same person. We must note that Lord and Novick's idea has had criticisms as their idea was a "thought experiment" and it can never be done (Borsboom (2005), Novick and Jackson (1974)). In this chapter, we briefly overview important topics useful in the rest of the dissertation.

1.1 Misclassification

Misclassification can be considered a special case of measurement error specifically for the situation when measurement is the categorical classification of events. Misclassification occurs when subjects incorrectly report their status or participation in a particular program. There is substantial evidence that misclassification is at least somewhat prevalent in a variety of situations in which individuals self-report (Oliveira et al. (2009), Pérez-Stable et al. (1992)). Its prevalence within nearly every field of statistics is a by-product of life in an imperfect world, and statistical inference that ignores misclassification introduces bias into the estimation and decision-making process.

The general approach to estimation in the presence of misclassified data can be summarized by the following steps. First, we must assume that the true classification status for an observation exists, which we may denote with the random variable Y where individuals belong to group $Y = 1$ with probability τ and to group $Y = 0$ with probability $(1 - \tau)$. In most cases we will asuume that direct observation of Y is impossible. There are notable exceptions where the ability to observe Y does exist, although the means by which it is obtained may be prohibitively invasive or expensive. In these situations, cheaper and easier alternatives are desirable to use in conjunction with or replacement of Y. Furthermore, we assume that the measurement process or processes observe scalar or vector T, which we assume has some relationship to the true exposure status Y. In the case of binary scalar T, we define the sensitivity as $Se = pr(T = 1|Y = 1)$, and the specificity as $Sp = pr(T = 0|Y = 0)$. Conversely, we may choose to characterize the relationship between T and Y in terms of misclassification rates, where we define the false negative rate $\theta_- = pr(T = 0|Y = 1)$ and the false positive rate $\theta_+ = pr(T = 1|Y = 0)$. However, we note that $Se = 1 - \theta_-$ and $Sp = 1 - \theta_+$, and thus either approach can lead to the proper adjustment of the resulting estimates.

A Bayesian approach to misclassified or mismeasured data has some important advantages over the frequentist approach. The use of priors with misclassified data help narrow the bounds on unidentified coefficients relative to the bounds estimated in a frequentist regression context. Also note, the Bayesian approach more easily incorporates various parameter constraints, such as restrictions on the extent of misclassification.

1.2 Bayesian Estimation

Statistical inference based on "classical" or "frequentist" methods places a distributional assumption on a random variable or vector represented by X with

support $\mathcal{X}$ that is assumed to be governed by fixed parameter vector θ. Because θ is typically unknown, the goal of inference is to observe some randomly selected sample $\mathbf{x}' = (x_1, \ldots, x_n)$ through which estimates of θ are empirically derived. The most utilized approach generally involves maximizing the likelihood function $L(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta)$ where $f(\cdot)$ is the probability density function of X.

The Bayesian approach to statistical inference differs from the frequentist approach such that θ is assumed to be a random variable rather than a fixed quantity. As a random variable, we define θ to have range Θ as well as its own density function $p(\theta)$. We refer to this function as the prior distribution which like any other density function contains all known information about θ. However, practitioners frequently lack prior knowledge of a parameter. This often leads to diffuse priors which implies that the prior variance is quite large. For more information on the elicitation and use of prior distributions, see Robert (2001).

Inference from a Bayesian perspective thus combines our prior information of the parameter with the observed knowledge gained from the likelihood using Bayes Theorem to yield a posterior distribution

$$p(\theta|\mathbf{x}) = \frac{p(\theta) f\left(\mathbf{X}|\theta\right)}{\int_{\theta\in\Theta} p(\theta) f(\mathbf{x}|\theta) d\theta}.$$

All posterior information on θ is contained within $p(\theta|\mathbf{x})$, which may or may not have a closed form. For a far more thorough and meaningful introduction to Bayesian inference see Lee (2004), Gelman et al. (2004), and Robert (2001).

Historically, the limitations to Bayesian inference dealt primarily with the often complicated and intractable form of the posterior distribution $p(\theta|\mathbf{x})$. That is, the posterior distribution is often intractable due to difficulties in estimating the denominator of Bayes rule. However, the advent of Markov Chain Monte Carlo (MCMC) methods have allowed practitioners of Bayesian inference to obtain numerical representations of the posterior distribution via computationally intensive

tools such as the Gibbs Sampler and Metropolis-Hastings algorithm. A thorough treatment of MCMC methods can be found in Robert and Casella (2004).

The Bayesian paradigm for estimation of misclassified binary data simply behaves as a missing data problem in which Y is unobserved for some or all individuals, and posterior estimates of Y and the related parameter τ can be produced via imputation of the missing observations from a Markov Chain Monte Carlo sampler. Thus, imputation of Y and estimation of τ are a part of the same process. For more information on this convenient consequence of Bayesian estimation see Carroll et al. (2006).

1.3 Modeling Probabilities Associated with Binary Outcomes

Modeling data with binary outcomes typically utilizes three link functions to relate the response probability $\pi_i = pr(X_i = 1)$ to the covariate vector z_i:

(1) the logit link:

$$f(\pi_i|\mathbf{z}_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i\beta.$$

(2) the probit link:

$$f(\pi_i|\mathbf{z}_i) = \mathbf{\Pi}^{-1}(\pi_i) = \mathbf{z}_i\beta$$

where $\mathbf{\Pi}^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution.

(3) the log link:

$$f(\pi_i|\mathbf{z}_i) = \log(\pi_i) = \mathbf{z}_i\beta.$$

We will proceed using the logit link function in Chapter 2 due to its ease of interpretation of model parameters. The estimated value of parameter β_k associated with covariate z_k is interpreted as the change in the log-odds of the response per unit change in z_k, where the odds are defined as $\pi/(1-\pi)$. The interpretation is not quite as straightforward using alternative link functions. Agresti (2002) Chapters

4 and 5 provides more information on modeling discrete outcomes and the use of logistic regression.

1.4 Bayesian Sample Size Determination

Because of the high cost and time constraints for clinical trials, researchers often need to determine the smallest sample size that provides accurate inferences for a parameter of interest. Although most experimenters have employed frequentist sample-size determination methods, the Bayesian paradigm offers a wide variety of sample-size determination methodologies. Bayesian sample-size determination methods are becoming increasingly more popular in clinical trials because of their flexibility and ease of interpretation.

Considerable attention has been given to Bayesian approaches to sample size determination. Dendukuri et al. (2004) consider several interval estimation criteria for the one sample binomial case. Branscum et al. (2007) consider a hypothesis testing criterion for the sample size problem for estimating sensitivity and specificity in a hierarchical model with multiple sites. Cheng et al. (2009) also apply a hypothesis testing criterion for a binary regression model when the outcome is subject to misclassification.

1.5 Problem of Underreported Data

In many count data applications, the recorded counts may only be a fraction of the true counts. Failing to adjust for such underreporting can lead to incorrect estimates of model parameters. Winkleman (1996) has addressed this problem by developing a Poisson regression model to adjust for underreported data using a Bayesian approach. His model uses a mixture of Poisson and binomial distributions to model the counts. The underreporting problem has also been considered by Ramos (1999) and Fader and Hardie (2000) who have used models based on the Winkleman approach. Ramos used the model to adjust for underreported counts of port wine

purchases for households in Portugal. Fader and Hardie have also analyzed the port wine data, but they use a beta-binomial/negative binomial distribution to model the counts. Stamey et al. (2004) used the same prior structure as Fader and Hardie and modeled underreported data to estimate sample size in a Bayesian setting.

1.6 Meta Analysis

Meta anlaysis is a statistical technique for combining the findings from independent studies. A meta analysis is most often used to assess the clinical effectiveness of healthcare interventions. Precise estimates of treatment effect can be provided, giving due weight to the size of the different studies included. It investigates not only the reported results of the studies but all aspects of research designs that produced them, including theoretical constructs, operational definitions of the independent variable, population samples, data collection procedures, statistical analysis, and especially the handling of possible confounding variables that would provide an alternative explanation for the reported results.

In the context of research about the outcomes of interventions to preserve or enhance physical, psychological, or social functioning, meta-analysis addresses two principal questions:

(1) Is there support in the sampled population of studies for the causal inference that the intervention made a statistically significant difference in the outcome(s)? And if so,

(2) how large an effect or difference did the intervention make?

Meta analysis results are commonly displayed graphically as forest plots. Forest plots visually display the effect size of all the studies, and the results of the meta analysis. Cochran's Q test is a statistical test used in conjunction with the forest plot to determine the significance of heterogeneity among studies. An alternative approach that quantifies the effect of heterogeneity is the I^2 statistic. This quantitiy

describes the percentage of total variation across studies that is due to heterogeneity rather than chance.

1.7 Organization

The chapters are divided as follows: in Chapter 2 we introduce a Bayesian estimation method using logistic regression in the presence of misclassified covariates and response and apply the model to a data set, which includes students who claim to have a math learning disability and were put through a series of tests, including a psychological evaluation. A simulation was also performed to determine the strength and validity of our model. In Chapter 3 and 4 we discuss a sample size determination method for efficacy and safety in a clinical trial. In Chapter 3 the efficacy and safety variables are assumed independent. We allow the safety variable to be underreported, and compare the underreported model to the non-underreported model. In Chapter 4, we use a regresion model to allow the efficacy and safety variables to be dependent. Again, we account for underreporting in the safety variable and compare the results to the model not accounting for underreporting. In Chapter 5, we discuss a meta analysis project I worked on with Dr. James Edgerton, a heart surgeon at The Heart Hospital Baylor Plano. We provide cumulative results of a preliminary meta analysis of the available stand alone atrial fibrillation (AF) surgical intervention publications from 2009 to 2011. Finally, in the appendices we include the R and WinBUGS code for execution of the methods described in the chapters.

CHAPTER TWO

Bayesian Estimation of Logistic Regression with Misclassified Covariates and Response for Educational Psychology Data

During the last decade, measurement error problems in binary regression are found to be of considerable interest among the researchers. The importance is particularly felt in analyzing data arising out of epidemiologic studies, in quantal bioassay problems and in many other important areas where the responses are binary in nature (Roy et al. (2005)). In epidemiologic studies, a serious source of error is misclassification of binary responses.

Misclassification can be considered a special case of measurement error specifically for the situation when measurement is the categorical classification of items. Its prevalence within nearly every field of statistics is a by-product of life in an imperfect world, and statistical inference that ignores misclassification introduces bias into the estimation and decision-making process.

Frequentist methods, or "classical" methods, place a distributional assumption on a random variable or vector, represented by Y with support $\mathcal{Y}$, that is assumed to be governed by fixed parameter vector θ. Because θ is typically unknown, the goal of inference is to observe some randomly selected sample $\mathbf{y} = (y_1, \ldots, y_n)$ through which estimates of θ are empirically derived. The most utilized approach generally involves maximizing the likelihood function $L(\theta|y_1, \ldots, y_n) = f(y_1, \ldots, y_n|\theta)$ where $f(\cdot)$ is the probability density function of Y.

The Bayesian approach to statistical inference differs from the frequentist approach in the sense that θ is assumed to be a random variable rather than a fixed quantity. As a random variable, we define θ to have range Θ as well as its own density function $p(\theta)$. We refer to this function as the prior distribution which like

any other density function contains all known information about θ. Inference from a Bayesian perspective thus combines our prior information of the parameter with the observed knowledge gained from the likelihood using Bayes Theorem to yield a posterior distribution

$$pr(\theta|\mathbf{y}) = \frac{pr(\theta)f(\mathbf{Y}|\theta)}{\int\limits_{\theta\in\Theta} pr(\theta)f(\mathbf{Y}|\theta)d\theta}.$$

All posterior information on θ is contained within $pr(\theta|y)$, which may or may not have a closed form.

Bayesian methods offer practical advantages over frequentist methods for the analysis of epidemiologic data by allowing the possibility of incorporating relevant prior scientific information and the ability to make inferences that do not rely on large sample assumptions. In this paper, we consider a fully Bayesian analysis that affords such adjustments, accounting for the sources of error and correcting estimates of the regression parameters. Unlike exisiting methods, our approach does not need to assume any parameters are known.

Section 2.1 describes the methods and priors used. We describe the data in Section 2.2, and the results of applying our models are given in Section 2.3. We conclude with a discussion.

2.1 Statistical Methods

Suppose that a diagnostic test is available to detect the presence or absence of a certain condition, D. Let Y designate the true disease status, with $Y+$ if D is present and $Y-$ otherwise. Similarly, let $T+$ and $T-$ denote test positive and test negative outcomes on a given dichotomous test.

In the Bayesian paradigm, we combine our prior information of the parameter with the observed knowledge gained from the likelihood using Bayes rule to yield a posterior distribution

$$pr(\theta|\mathbf{x}) = \frac{pr(\theta)f\left(\mathbf{X}|\theta\right)}{\int_{\theta\in\Theta} pr(\theta)f(\mathbf{x}|\theta)d\theta}.$$

Estimating the prevalence of a disease, θ, will depend on the sensitivity $Se = pr(T+|Y+)$ and specificity $Sp = pr(T-|Y-)$, where $pr(A|B)$ is the conditional probability of A given B, which can be represented as

$$pr(A|B) = \frac{pr(A \cap B)}{pr(B)}.$$

Since Se and Sp for each test are not exactly known, they have to be estimated along with prevalence, θ. Table 2.1 reviews how estimates are constructed from the model with the two dichotomous tests. For each subject, there are two available tests to detect a disease. Each test has two possible outcomes: positive or negative. This leads to four possible combinations for the observed data. If we consider the true disease status of each individual, we have eight possible combinations of observed and latent data. Let $Z_1, \ldots, Z_4$ be the latent data that represents the number of subjects with a positive disease status out of $a, \ldots, d$ subjects in each possible category for the observed test result, respectively. With the results from two dichotomous tests, one can calculate the probability of having a disease.

Table 2.1. Likelihood Contributions of All Possible Outcomes

Truth	Test 1 result	Test 2 result	Likelihood contribution per subject	No. of subjects
$+$	$+$	$+$	$\theta Se_1 Se_2$	Z_1
$+$	$+$	$-$	$\theta Se_1(1-Se_2)$	Z_2
$+$	$-$	$+$	$\theta(1-Se_1)Se_2$	Z_3
$+$	$-$	$-$	$\theta(1-Se_1)(1-Se_2)$	Z_4
$-$	$+$	$+$	$(1-\theta)(1-Sp_1)(1-Sp_2)$	$a-Z_1$
$-$	$+$	$-$	$(1-\theta)(1-Sp_1)Sp_2$	$b-Z_2$
$-$	$-$	$+$	$(1-\theta)Sp_1(1-Sp_2)$	$c-Z_3$
$-$	$-$	$-$	$(1-\theta)Sp_1Sp_2$	$d-Z_4$

To determine the probability of a true positive disease status, we take the likelihood contribution for the appropriate test results divided by the sum of the likelihood contributions for a true positive and true negative disease status with

the same dichotomous test results. For example, if a subject is positive on both dichotomous tests, then using rows 1 and 5 from Table 2.1, we reach the following probability:

$$pr(\text{subject is a true positive}) = pr(Y+) = \frac{\theta Se_1 Se_2}{\theta Se_1 Se_2 + (1-\theta)(1-Sp_1)(1-Sp_2)}.$$

We note that Table 2.1 can be expanded or reduced based on the number of dichotomous tests available.

The full likelihood function of the observed, W, and latent, Z, data is proportional to the product of each entry in the 'likelihood contribution' column of Table 2.1 raised to the power of the corresponding entry in the 'number of subjects' column of the table. For the case of two tests we have:

$$\begin{aligned} f(T_1, T_2, Z|\theta, Se_i, Sp_i) \propto & \\ & \prod [\theta Se_1 Se_2]^{z_i} [\theta Se_1(1-Se_2)]^{z_2} [\theta(1-Se_1)Se_2]^{z_3} \\ & \times [\theta(1-Se_1)(1-Se_2)]^{z_4} [(1-\theta)(1-Sp_1)(1-Sp_2)]^{a-z_1} \\ & \times [(1-\theta)(1-Sp_1)Sp_2]^{b-z_2} [(1-\theta)Sp_1(1-Sp_2)]^{c-z_3} \\ & \times [(1-\theta)Sp_1 Sp_2]^{d-z_4} \end{aligned} \tag{2.1}$$

where $T_1 = (T_{11}, \ldots, T_{1n})$ and $T_2 = (T_{21}, \ldots, T_{2n})$ are the results for the two dichotomous tests across all n subjects, respectively, and Se_i and Sp_i are the sensitivities and specificities associated with each test.

Estimates are derived either by maximum likelihood methods or by Bayesian methods with the addition of prior distributions for the prevalence θ and the test parameters, Se_i and Sp_i, $i = 1, 2$.

2.1.1 Continuous Test

Alternatively, one may encounter a continuous test, instead of a dichotomous test. We assume the continuous test, for disease prevalence, follows normal distribution for each sub-population. Thus, $N(\mu_D, \sigma_D^2)$ is the density function for the results

of the continuous test conditional on having the disease, and $N(\mu_{ND}, \sigma^2_{ND})$ is the density function for the results of the continuous test conditional on not having the disease. The likelihood for the observed and latent data is

$$f(W, Z|\theta, \mu_D, \sigma^2_D, \mu_{ND}, \sigma^2_{ND}) = \prod_{i=1}^{n} \left(\theta \frac{1}{\sqrt{2\pi}\sigma_D} \exp\left\{ -\frac{1}{2\sigma^2_D}(w_i - \mu_D)^2 \right\} \right)^{z_i} a \times \left((1-\theta) \frac{1}{\sqrt{2\pi}\sigma_{ND}} \exp\left\{ -\frac{1}{2\sigma^2_{ND}}(w_i - \mu_{ND})^2 \right\} \right)^{1-z_i} \quad (2.2)$$

where W are the observed continuous test values, and Z are the latent data. Thus, the likelihood contribution from each subject i is a normal distribution.

In addition to the continuous test, we consider the case where we have two dichotomous tests from above. The full likelihood function for data from all three methods is a combination of the above likelihood (2.2) with that from two dichotomous tests (2.1) implied by Table 2.1. The full likelihood function for the outcome model is

$$\begin{aligned} f(W, Z, T_1, T_2|\theta, Se_1, Se_2, Sp_1, Sp_2, \mu_D, \sigma^2_D, \mu_{ND}, \sigma^2_{ND}) = \\ \prod_{i=1}^{n} \Big(\theta Se_1^{T_{1i}} (1-Se_1)^{(1-T_{1i})} Se_2^{T_{2i}} (1-Se_2)^{(1-T_{2i})} \\ \times \frac{1}{\sqrt{2\pi}\sigma_D} \exp\left\{ -\frac{1}{2\sigma^2_D}(w_i - \mu_D)^2 \right\} \Big)^{z_i} \\ \times \Big((1-\theta)(1-Sp_1)^{T_{1i}} Sp_1^{(1-T_{1i})} Sp_2^{(1-T_{2i})} (1-Sp_2)^{(1-T_{2i})} \\ \times \frac{1}{\sqrt{2\pi}\sigma^2_{ND}} \exp\left\{ -\frac{1}{2\sigma^2_{ND}}(w_i - \mu_{ND})^2 \right\} \Big)^{1-z_i} \end{aligned} \quad (2.3)$$

where $T_1 = (T_{11}, \ldots, T_{1n})$ and $T_2 = (T_{21}, \ldots, T_{2n})$ are the results for the two dichotomous tests across all n subjects, respectively.

2.2 Covariate Misclassification

Categorical covariates are often subject to misclassification, and this misclassification can distort the relationship of the outcome of interest. Failing to account

for misclassification yields biased and inconsistent coefficient estimates. To account for this error, we assume the true status of the covariate E, in the exposure model, is

$$f(E|p_E) = p_E^E(1-p_E)^{1-E}, \tag{2.4}$$

where p_E is based on expert opinion or prior beliefs, and E is often the exposure variable. This model can also be expanded to include other covariates, such that

$$\text{logit}(p_E) = \gamma_0 + \gamma_1 \mathbf{U},$$

where $\mathbf{U}$ is an additional measured confounder.

We must now consider the fallible diagnostic test for exposure in the measurement model

$$f(X|p_X) = p_X^X(1-p_X)^{1-X}, \tag{2.5}$$

where $p_X = ESe_X + (1-E)(1-Sp_X)$, and Se_X and Sp_X are the sensitivity and specificity, respectively, for the observed covariates of interest. Based on the above formulations, the full model is the product of (2.3), (2.4), and (2.5), thus

$$\begin{aligned} &f(W, Z, T_1, T_2, E, X|\Theta) \\ &\qquad = \prod_{i=1}^{n} f(W, Z, T_1, T_2|\theta, Se_1, Se_2, Sp_1, Sp_2, \mu_D, \sigma_D^2, \mu_{ND}, \sigma_{ND}^2) \\ &\qquad \times f(E|p_E) f(X|p_X) \end{aligned}$$

where Θ consists of the remaining parameters.

Rather than using fixed prior distributions, one can employ hierarchical modelling, which allows the investigation of the effects of any covariates on the prevalence θ or any other test properties, such as the sensitivities and specificities. This can be applied to the outcome probabilities. The α and β parameters can be directly modelled via hierarchical distributions, such as with the gamma distribution. However, we will follow a more common method using the logistic function.

We will replace the beta prior distribution for θ and let θ_i represent the probability of subject i having the disease, and model as follows:

$$\gamma_i = \text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right)$$

$$\gamma_i \sim N(\mu_{\gamma_i}, \sigma^2_\gamma)$$

$$\mu_{\gamma_i} = \beta_0 + \beta_j E_{ij}$$

where, E_{ij} are the covariates of interest for each subject i. The posterior distribution of $\exp(\beta)$, then estimate the odds ratios for the corresponding covariates. Normal prior distributions can be used for the regression parameters β_0 and β_1, and a uniform prior distribution can be used for σ_γ.

We must note that this model can be expanded or simplified depending on the number of tests, the type of test, and whether or not covariates are included in the model.

2.2.1 Prior Distributions

An important step in any Bayesian analysis is to obtain a prior distribution over all model parameters. This can be accomplished using past data, if available, or expert opinion. There is a large amount of literature on the elicitation of prior distributions. Proposed methods have included directly matching percentiles (Press (1989)) or means and standard deviations (Lee (2004)) to a member of a preselected family of distributions of the data (Chaloner and Duncan (1983)). The predictive distribution is the marginal distribution of the observable data, which is found by integrating the likelihood of the data over the prior distribution of the unknown parameters (Lee (2004)).

Sensitivity and specificity of diagnostic tests are generally not known exactly, though probability models for such quantities are often available, based either on previous data or expert opinion. A beta distribution is a natural choice for modelling uncertainty about a probability. To transform expert opinion into a particular beta

distribution, it suffices to elicit the prior mode and the quantity associated with the 95th (or some other) percentile. Beta distributions satisfying such requirements are easily determined and subsequently graphically examined by the expert. A free GUI-based statistical package called BetaBuster can be used to obtain specific Beta prior distributions based on scientific input. The choice of a Be(a, b) distribution can be made by specifying the parameters a and b. We assume the prior for Se is Be(a_{Se}, b_{Se}), independent of the Be(a_{Sp}, b_{Sp}) prior for Sp.

In our model, we will look at two sets of priors for the sensitivities and specificities. First, we will center our priors around the true value of the sensitivities and specificities. Our second set of priors will be slightly offset from the true values. Along with the two sets of priors will we look at two different values for the truth for each sensitivity and specificity. This will give us four models total to follow.

For the exposure model, which incorporates the true status of the covariates within the model, we give p_E a Beta distribution. This prior, as per the rest, is based on expert opinion.

A common prior for β_0 and β_1 would be independent normal distributions with zero mean and a common large variance. While a simple 'non-informative' choice is appealing, we would argue that some care is necessary. If the variance is too large, the induced prior on β_0 or β_1 could cause problems in our analysis. Thus, some care is necessary in the choice of the variance in order to obtain a reasonable 'non-informative' normal prior for the regression coefficients. We will discuss this in more detail in Section 2.4.4.

2.2.2 Markov Chain Monte Carlo (MCMC) Implementation

In the analysis of our data, we used a flexible software for the Bayesian analysis of complex statistical models using MCMC methods. This iterative algorithm builds approximations to the posterior distributions of interest based on Monte Carlo simu-

lations. Compared with alternatives such as numerical integration, MCMC requires much less computing time and is easier to implement and customize.

The WinBUGS software was used to carry out the MCMC simulations from the posterior distribution, and the code is available in Appendix A. The results are based on 5000 iterations, after a burn-in of 1000. To reduce autocorrelation within the chains, we retained only every third iteration, a process known as thinning. The samples from every kth iteration will be used for inference, where k is the thinning value. Setting $k > 1$ can help to reduce the autocorrelation in the sample.

2.3 The Data

2.3.1 Participants

The Missouri Math Difficulties Questionnaire (MMDQ) was designed as a 21-item instrument, utilizing a 5-point Likert-type scale that the respondent would complete with an evaluator as part of a more comprehensive intake evaluation. Each item was worded so as to ascertain how much difficulty the respondent had doing a "real world" task that involves some aspect of math, such as balancing his/her check book and calculating tips.

The MMDQ was given to 546 college students from two different universities as part of a larger study of college students with math disorders. The sample from university one contained 159 students who participated as part of a research requirement in an Introduction to Psychology class, none of which indicated having any diagnosed math disorders. The sample from university two contained 387 students who were self referred to an assessment clinic because of psychoeducational difficulties. They received a complete psychoeducational evaluation that included the MMDQ as well as full cognitive, academic, memory, personality, and attention assessments. A math disorder diagnosis was based on the judgement of a team which included a licensed psychologist.

Participation in the study was solicited from two separate projects (Project One and Project Two) at a large midwestern university. The participants in Project One (n = 93) were undergraduate students who referred themselves to a university-based psychology clinic due to having self-identified math difficulties. Seventy-nine of these students were part of a research project designed to study technology with college students with a math-related learning disability, and fourteen of the students were assessed as part of the clinic's regular clientele. All students had difficulties completing a university-wide required algebra class (i.e., failed the required college algebra course at least once or passed the course with significant difficulties).

Project Two involved a research study at the same midwestern university, designed specifically to gather the same data as gathered in Project One from a community control group (Kazdin (2003)). All Project Two participants had to meet the following criteria: (a) have undergraduate level status; (b) have no self-reported math difficulties; (c) no major in math or a math-related field (e.g., statistics); and (d) have no current or previous Diagnostic and Statistical Manual of Mental Disorders - Fouth Edition - Text Revision (DSM-IV-TR) diagnoses (APA (2000)).

2.3.1.1 *Instrument: Woodcock Johnson – Third Edition Tests of Achievement.* The Woodcock Johnson - Third Edition (WJ-III) Tests of Achievement (Woodcock et al. (2001)) is a standardized battery of individually administered academic achievement tests. There are three math subtests in the core battery: (a) Applied Problems (AP; i.e., ability to orally answer spoken math word problems); (b) Math Fluency (MF; i.e., ability to answer single-digit addition and subtraction problems within three minutes); and (c) Calculation (C; i.e., ability to compute answers to math problems ranging from simple addition to calculus without time restrictions,). McGrew and Woodcock (2001) estimated the internal consistency of the three Woodcock Johnson - Third Edition Tests math scores to be between 0.88 - 0.89 for Applied

Problems, between 0.82 - 0.85 for Math Fluency, and between 0.92 - 0.93 for Calculation, for 18-29 year-olds.

2.4 Results

2.4.1 Simulated Data

In this section, we investigate the performance of the proposed algorithm. We use a simulated data set and test the performance under a number of alternative specifications of the data generating process. We use WinBUGS to run the MCMC simulations for a sample size of $n = 1000$. We simulate the data in R before sending it to WinBUGS, and we repeat this process 500 times. Our results below are from averaging the outcomes from the 500 repetitions.

We will look at four different simulation results. We varied the true value of the sensitivities and specificities and we allowed the priors to be centered and offset. In the first simulation, we have a true sensitivity of 0.9 and a true specificity of 0.7, and the priors are centered at the true values. The results are given in Table 2.3. The table displays the true value of each parameter along with the mean value from multiple repetitions (500) through WinBUGS. Sensitivity and specificity are given for Y, the true disease status, and Se_X and Sp_X are given for the exposure variable.

For the simulated data sets, we have the following priors:

Table 2.2. Priors for the Simulated Data

Variable	Prior
Se (centered)	Beta(90, 10)
Se (offset)	Beta(84, 16)
Sp (centered)	Beta(70, 30)
Sp (offset)	Beta(76, 24)
β_0	Beta(0, 0.1)
β_1	Beta(0, 0.1)
Se_X	Beta(80, 20)
Sp_X	Beta(90, 10)

The table also displays the standard deviation and 95% credible interval. The standard deviations for Se_X and Sp_X are larger than those for Se and Sp. We are relying on our prior only, not the data, and since the priors are fairly wide, we have high coverage. As an extension to this model, we could add a second diagnostic test for X, the fallible diagnostic test for exposure. This will tighten the intervals and bring the coverage closer to nominal. The coverage for each estimate, displayed in the last column, is the percentage of time the mean value was captured in the credible interval for each iteration through WinBUGS.

The final row in the table displays the deviance, which evaluates the goodness of fit of our model. The definition of deviance is $-2 \times \log(\text{likelihood})$: 'likelihood' is defined as $p(y|\theta)$, where y comprises all stochastic nodes given values (i.e. data), and θ comprises the stochastic parents of y ('stochastic parents' are the stochastic nodes upon which the distribution of y depends) when collapsing over all logical relationships.

Table 2.3 provides the results for the first simulation. Our model accurately estimated each parameter of interest. Our credible intervals are of reasonable width for the model. In the first simulation, the coverage for sensitivity, specificity, and our β's were all high. As explained before, Se_X and Sp_X have a coverage of 1 because we are relying on only the prior information, not the data.

Table 2.3. True Sensitivity of 0.9 and True Specificity of 0.7, With Centered Priors

Parameter	actual	mean	sd	2.5%	97.5%	coverage
Se	0.9	0.901	0.011	0.878	0.921	0.968
Sp	0.7	0.700	0.021	0.658	0.741	0.976
β_0	-0.20	-0.216	0.141	-0.517	0.038	0.98
β_1	1.8	1.872	0.312	1.360	2.563	0.97
Se_X	0.8	0.797	0.036	0.725	0.864	1
Sp_X	0.9	0.897	0.029	0.834	0.947	1
deviance		12829.926	71.122	12682.704	12959.965	

Table 2.4 provides the results for the average of 500 repetitions through WinBUGS. In the second simulation, we have a true sensitivity of 0.7, a true specificity of 0.9, and, again, the priors are centered around the true values.

Table 2.4. True Sensitivity of 0.7 and True Specificity of 0.9, With Centered Priors

Parameter	actual	mean	sd	2.5%	97.5%	coverage
Se	0.7	0.698	0.017	0.665	0.731	0.97
Sp	0.9	0.901	0.014	0.872	0.926	0.958
β_0	-0.20	-0.216	0.142	-0.519	0.037	0.978
β_1	1.8	1.872	0.312	1.361	2.567	0.976
Se_X	0.8	0.798	0.036	0.725	0.864	1
Sp_X	0.9	0.897	0.029	0.834	0.947	1
deviance		12993.359	71.105	12846.151	13123.326	

The next two tables display the same information as above, for the remaining simulations. Simulation three, again averages over 500 iterations, for a sample size of $n = 1000$. In this simulation we have a true sensitivity of 0.9, a true specificity of 0.7, and the priors are slightly offset from the true value. The results are given in Table 2.5. Our estimates and credible intervals are very similar to the results in the first simulation. For the results from the offset priors, our coverage is less than the results from the centered priors in Table 2.3. However, we still have fairly high coverage.

Table 2.5. True Sensitivity of 0.9 and True and Specificity of 0.7, With Offset Priors

Parameter	actual	mean	sd	2.5%	97.5%	coverage
Se	0.9	0.892	0.011	0.869	0.913	0.92
Sp	0.7	0.711	0.021	0.669	0.751	0.926
β_0	-0.20	-0.219	0.142	-0.522	0.034	0.974
β_1	1.8	1.875	0.313	1.364	2.570	0.972
Se_X	0.8	0.797	0.036	0.725	0.864	1
Sp_X	0.9	0.897	0.029	0.834	0.947	1
deviance		12832.620	71.0674	12685.218	12962.746	

For the final simulation, we have the true sensitivity at 0.7, the true specificity at 0.9, and the priors are slightly offset from the true value. The results are given in Table 2.6. Again, our estimates and credible intervals are very similar to the results in the second simulation. As with the previous simulation, the coverage is less than in Table 2.4, but the coverage is still adequate.

Table 2.6. True Sensitivity of 0.7 and True Specificity of 0.9, With Offset Priors

Parameter	actual	mean	sd	2.5%	97.5%	coverage
Se	0.7	0.708	0.017	0.674	0.740	0.938
Sp	0.9	0.888	0.015	0.858	0.915	0.886
β_0	-0.20	-0.210	0.141	-0.509	0.044	0.978
β_1	1.8	1.858	0.308	1.349	2.544	0.984
Se_X	0.8	0.797	0.036	0.724	0.864	1
Sp_X	0.9	0.897	0.029	0.835	0.947	1
deviance		12995.098	71.185	12847.651	13125.511	

The priors for our diagnostic tests are fairly wide, and we are relying on our prior information only. Thus, we have high coverage for Se_X and Sp_X. As an extension to this model, we could add a second diagnostic test for X, the fallible diagnostic test for exposure, as previously mentioned. This will tighten the intervals and bring the coverage closer to nominal.

2.4.2 Math Data

We applied our model to the math data set discussed in Section 2.3. Table 2.7 provides the priors for this model. We chose non-informative Beta priors for the regrssion coefficients and informative priors centered around the truth for the sensitivies and specificities. The regression coefficient β_1 refers to the exposure variable, this allows for the misclassification in our exposure variable. The remaining β's refer to each covariate - location of study (β_2), high school GPA (β_3), and gender (β_4). The Se and Sp estimates refer to the math variables. Se_1 and Sp_1 are the sensitivity and specificity for the math fluency variable, while Se_2 and Sp_2 are for the math

calculation skills variable. Finally, the Se_X and Sp_X estimates refer to the verbal variables - passage comprehension (Se_{X_1} and Sp_{X_1}) and reading fluency (Se_{X_2} and Sp_{X_2}). Table 2.8 displays the results, including the parameter estimates, standard deviations, and 95% credible intervals.

Table 2.7. Priors for the Math Data Set

Variable	Prior
Se_1	Beta(9, 1)
Sp_1	Beta(8, 2)
Se_2	Beta(9, 1)
Sp_2	Beta(8, 2)
β_0	Beta(0, 0.1)
β_1	Beta(0, 0.1)
β_2	Beta(0, 0.1)
β_3	Beta(0, 0.1)
β_4	Beta(0, 0.1)
Se_{X_1}	Beta(8, 2)
Sp_{X_1}	Beta(9, 1)
Se_{X_2}	Beta(8, 2)
Sp_{X_2}	Beta(9, 1)

As a comparison, we also analyzed the math data set without accounting for misclassification. These results are displayed in Table 2.9. Since we are not accounting for any misclassification, we are not interested in the sensitivities and specificities. For the results in which we do not account for misclassification, we arbitrarily chose to use Math Calculation Skills (MCS) and Passage Comprehension (PC) as the gold standard. For MCS, we used a cutoff of 80, meaning ≤ 80 indicates diagnosed and ¿ 80 indicates non-diagnosed. Likewise, for PC, we used a cutoff of 100, such that ≤ 100 represents diagnosed and ¿ 100 represents non-diagnosed. The coefficient β_1 represents the coefficient for Passage Comprehension, β_2 represents location of study, β_3 is for high school GPA, and β_4 is the coefficient for gender. Math Calculation Skills was used as our diagnostic test. The priors are the same as we used in the model accounting for misclassification.

Table 2.8. Math Data Accounting for Misclassification

Parameter	mean	sd	2.5%	97.5%
β_0	-0.7384	2.948	-6.536	4.952
β_1	-0.8169	2.88	-6.5	4.647
β_2	-1.849	1.726	-5.656	1.735
β_3	-0.8761	1.943	-5.962	1.882
β_4	0.5759	2.579	-5.223	5.076
Se_1	0.9045	0.08484	0.6874	0.9976
Sp_1	0.8941	0.02334	0.8438	0.9362
Se_2	0.9136	0.08017	0.7	0.9977
Sp_2	0.9579	0.01658	0.9218	0.9854
Se_{X_1}	0.357	0.03475	0.2912	0.4263
Sp_{X_1}	0.9008	0.08988	0.6619	0.9973
Se_{X_2}	0.4085	0.03539	0.3402	0.4798
Sp_{X_2}	0.8993	0.09158	0.662	0.9972

Table 2.9. Math Data Not Accounting for Misclassification

Parameter	mean	sd	2.5%	97.5%
β_0	-2.925	1.109	-5.131	-0.7291
β_1	0.2574	0.6774	-1.126	1.565
β_2	-0.2397	0.2958	-0.8343	0.3471
β_3	0.8965	0.2457	0.4287	1.39
β_4	0.6956	0.3052	0.08638	1.3

2.4.3 Convergence

Thinning is a common way to reduce autocorrelation in the sample in a simulation. We found that thinning of 3 removed autocorrelation and thus helped achieve convergence in the simulated data. The convergence plots below show the autocorrelation for the specified parameters in a given simulation.

2.4.3.1 *Simulated Data.* Figure 2.1 shows the history plots for β_0 and β_1 for the simulated data with sample size $n = 200$. This model did not require thinning or other remedial measures.

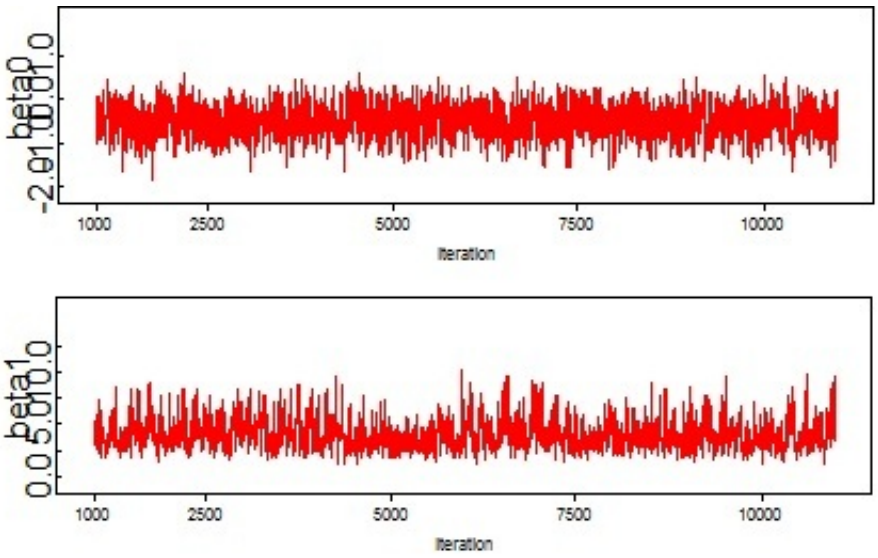

Figure 2.1: Convergence plots for β_0 and β_1 for our simulated data set with no thinning and a sample size of $n = 200$.

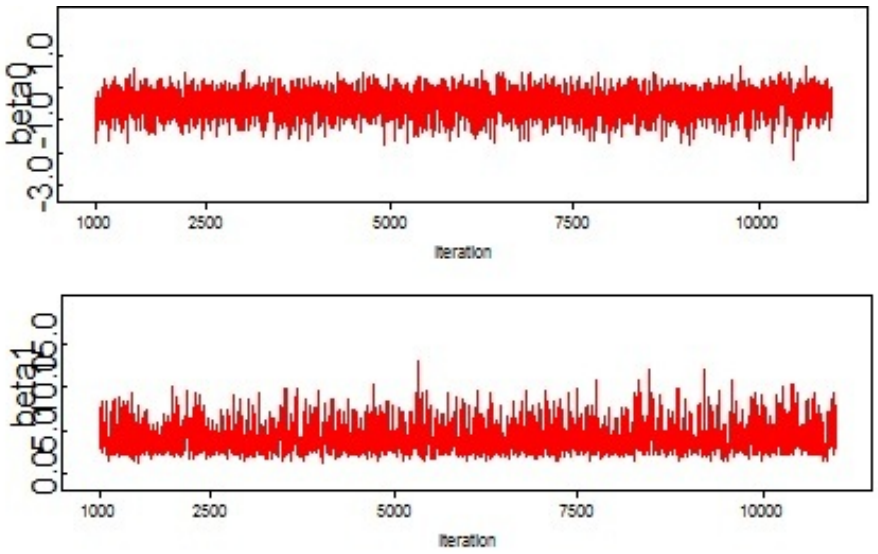

Figure 2.2: Convergence plots for β_0 and β_1 for our simulated data set with thinning of 3 and a sample size of $n = 200$.

Figure 2.2 shows the same autocorrelation plots of the β's for the simulated data with a sample size of $n = 200$. We noted that thinning, or some other measure, was necessary. Thus, we chose to thin at 3. We can see that after thinning we have a smoother autocorrelation plot, one that looks more like white noise.

For comparison purposes, we produces similar history plots for a larger sample size. Figure 2.3 shows the autocorrelation plots for β_0 and β_1 for the simulated data with a sample size of $n = 1000$. In this model we did not thin.

Figure 2.4 shows the same autocorrelation plots for the β's for the simulated data with a sample size of $n = 1000$. Again, some measure was necessary as Figure 2.3 does not look like white noise. We chose to thin at 3 for this example. We can see that as our sample size and thinning value increase, our autocorrelation plots smooth out and we produce history plots that look like white noise, which indicates no autocorrelation.

2.4.3.2 *Math Data.* Similar plots were constructed for the math data set. As with the simulated data, each data set included 10000 iterations after a 1000 iteration burn-in to produce the following results. Figure 2.5 displays the convergence plots of β_0, β_1, β_2, β_3, and β_4 for the math data set in which we account for misclassification. Some remedial measures are necessary, as the plots do not look like white noise. We chose different thinning rates until we reached our desired outcome. Figure 2.6 displays the history plots of the β's for the math data set not accounting for misclassification and a thinning rate of 10. Thinning produces much better history plots, though more measures may be necessary.

Finally, for the math data set in which we did not account for misclassification, Figure 2.7 displays the history plots of the β's. Thinning was not necessary as our plots look like white noise.

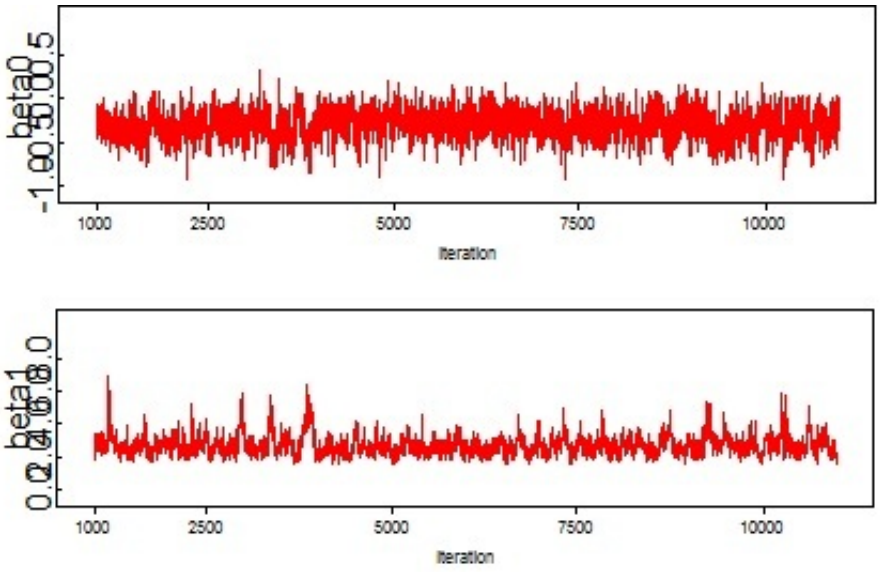

Figure 2.3: Convergence plots for β_0 and β_1 for our simulated data set with no thinning and a sample size of $n = 1000$.

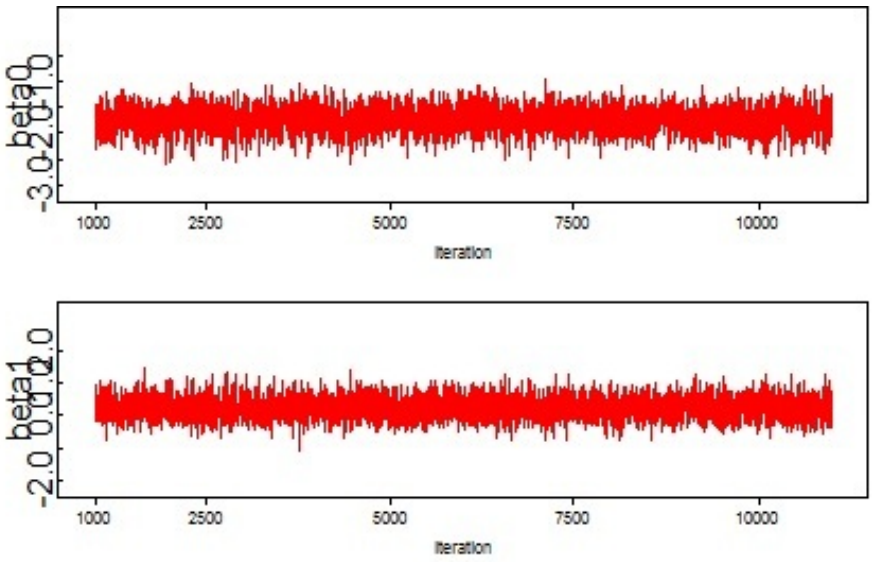

Figure 2.4: Convergence plots for β_0 and β_1 for our simulated data set with thinning of 3 and a sample size of $n = 1000$.

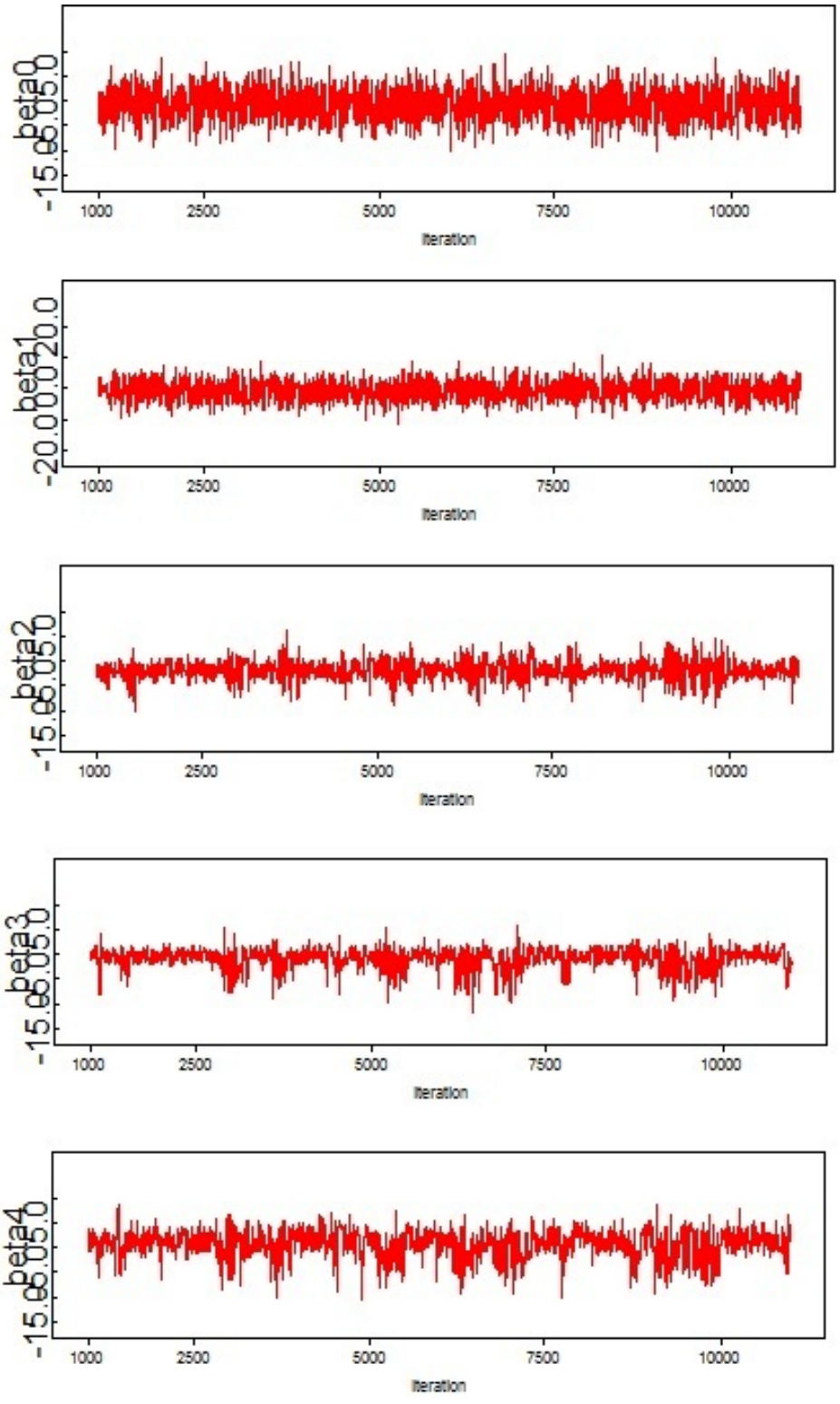

Figure 2.5: Convergence plots of the β's for the math data set that accounts for misclassification with no thinning.

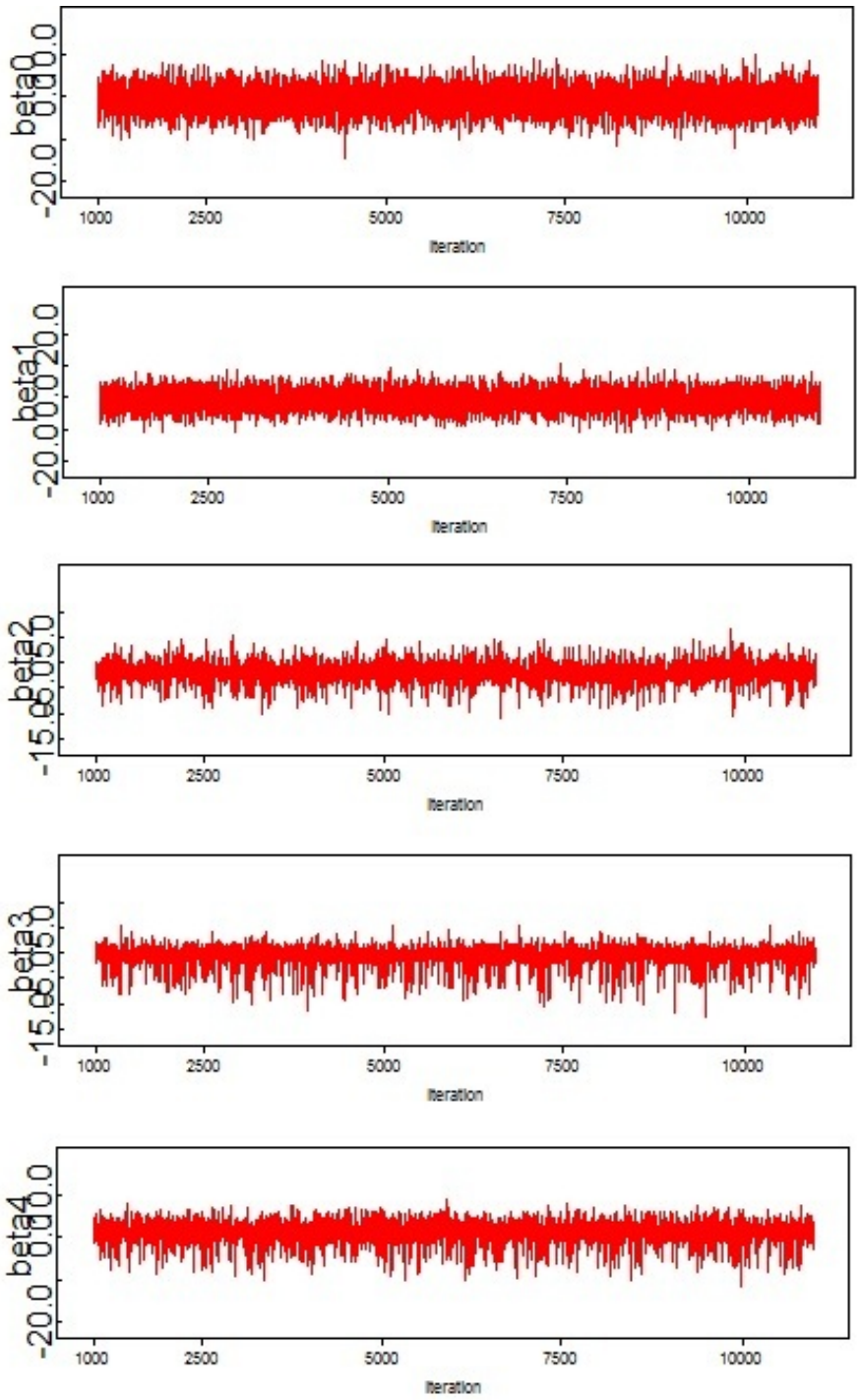

Figure 2.6: Convergence plots of the β's for the math data set that accounts for misclassification with thinning of 10.

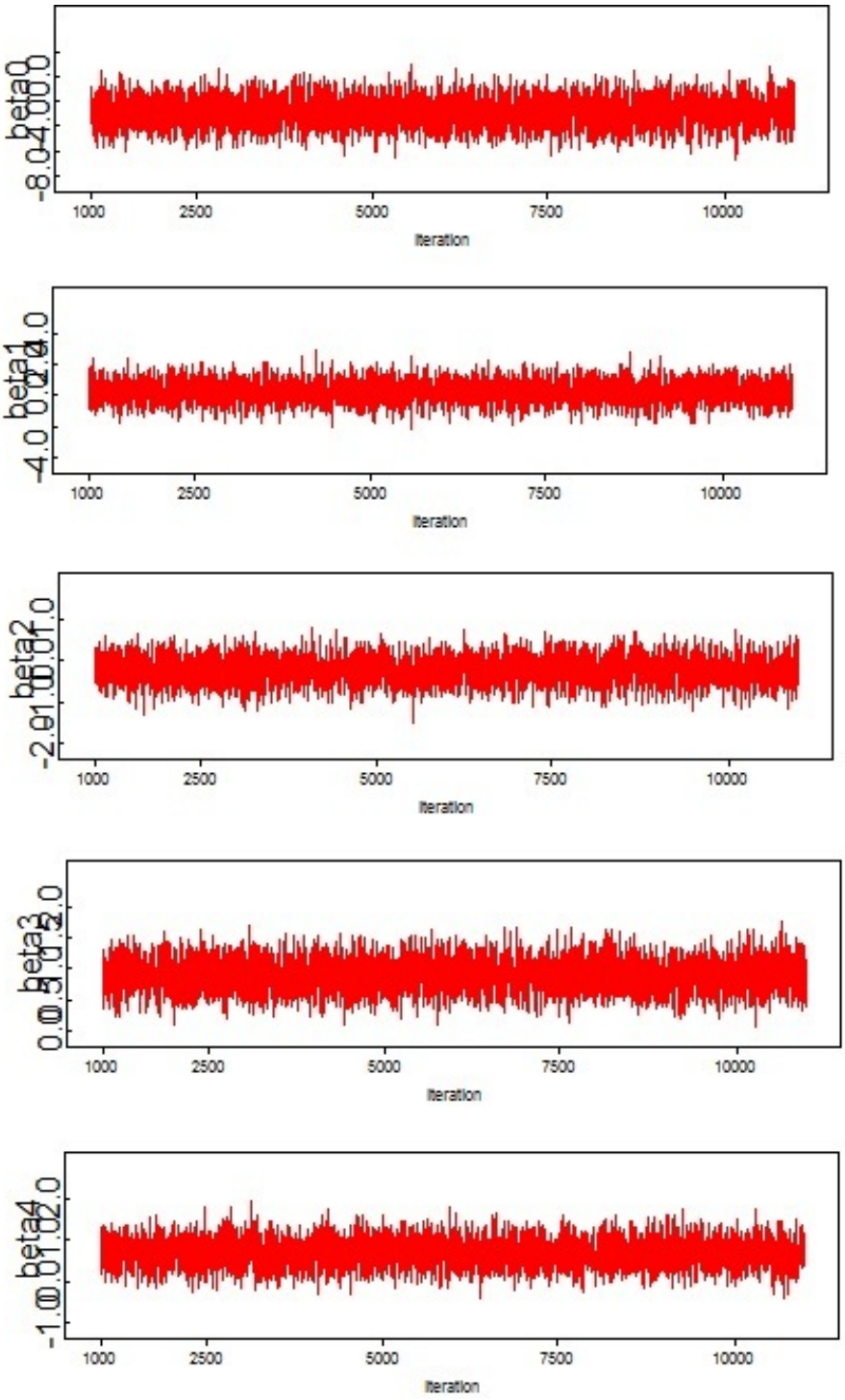

Figure 2.7: Convergence plots of the β's for the math data set that does not account for misclassification.

2.4.4 Varying Priors

In section 2.2.1 we discussed that care was necessary in choosing the prior variance for β. We will take a look at different prior variances to investigate the sensitivity to this selection the model exhibits.

We will look at four different priors distributions: highly diffuse, moderately difuse, moderately informative, and highly informative. For the highly diffuse priors, we use $\beta_0 \sim \text{Normal}(-2, 100)$ and $\beta_1 \sim \text{Normal}(0.5, 100)$. Figure 2.8 displays the history plots for β_0 and β_1 when the prior variances were highly diffuse. We can see that we have a lot of variability within the data, our parameter estimates are not as accurate. With highly diffuse priors, the parameter estimates for β_0 and β_1 are 3.7 and -7.8, respectively.

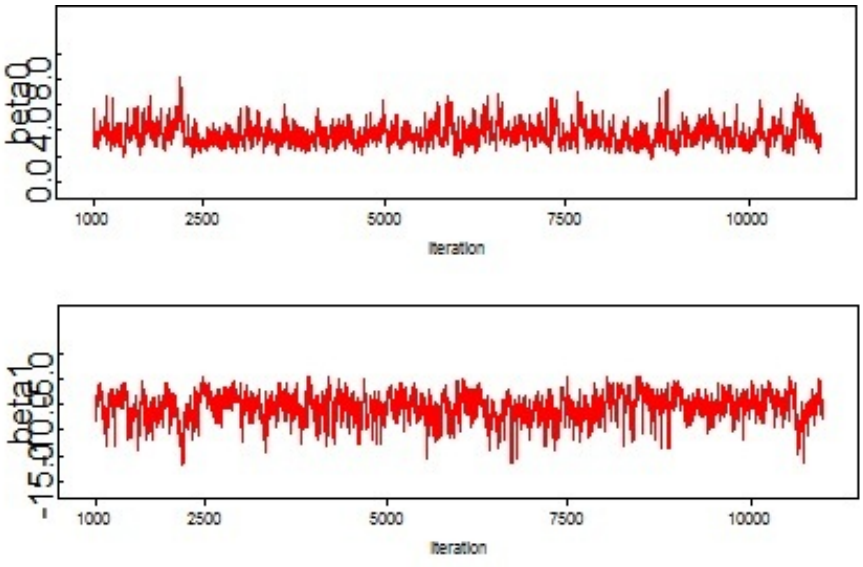

Figure 2.8: Convergence plots for β_0 and β_1 for one data set with highly diffuse prior variances.

For the moderately diffuse priors, we use $\beta_0 \sim \text{Normal}(-2, 25)$ and $\beta_1 \sim \text{Normal}(0.5, 25)$. The history plots are displayed in Figure 2.9. Again, the parameter estimates are not as accurate, and the history plots show much variability.

For the moderately informative priors, we use $\beta_0 \sim \text{Normal}(-2, 3)$ and $\beta_1 \sim \text{Normal}(0.5, 3)$. Figure 2.10 displays the history plots for β_0 and β_1 when the prior

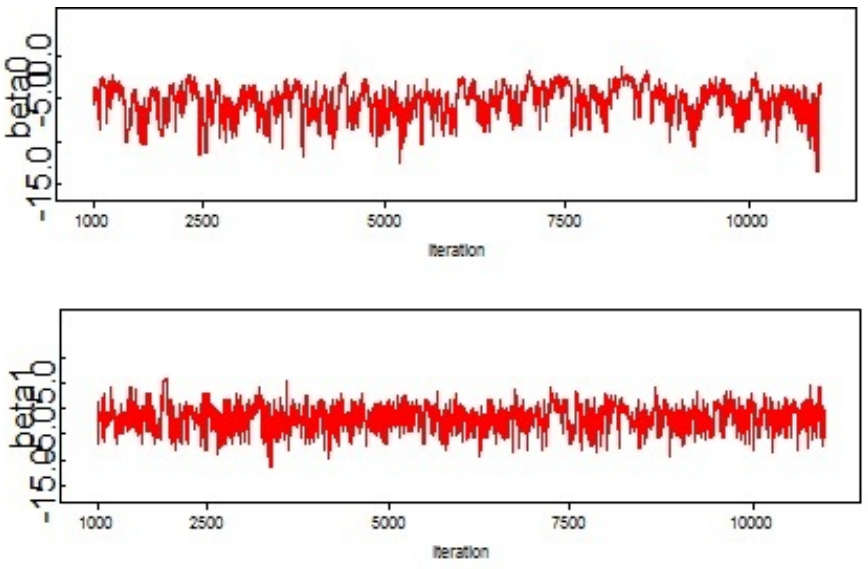

Figure 2.9: Convergence plots for β_0 and β_1 for one data set with moderately diffuse prior variances.

variances were moderately informative. We can see that we have less variability within the data compared to the previous two tables, and our parameter estimates are more accurate to the true value.

For the highly informative priors, we use $\beta_0 \sim \text{Normal}(-2, 1)$ and $\beta_1 \sim \text{Normal}(0.5, 0.5)$. The history plots are displayed in Figure 2.11. The parameter estimates are much more accurate with more informative priors. The parameters estimates for β_0 and β_1 are -2.15 and -0.3, respectively.

2.5 Discussion

The main advantage of a Bayesian approach for measurement error problems is that it allows the problem to be modeled in a conceptually straightforward way without approximations (Ren and Stone (2007)). All the available information is utilized and the uncertainty from different sources is properly reflected in the parameter estimates. Moreover, it works under more complicated model frameworks such as misclassification and measurement error.

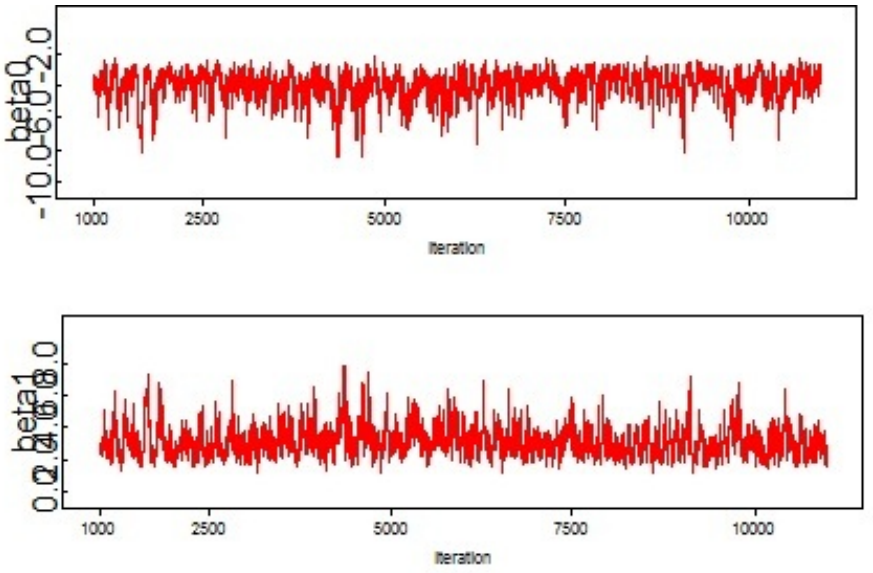

Figure 2.10: Convergence plots for β_0 and β_1 for one data set with moderately informative prior variances.

In this chapter, our methods incorporated the information from two dichotomous tests, one continuous test, and the misclassification of the covariates, using logistic regression, into a single model, while not considering any method to be a gold standard. We have improved on the previous work by combining all these tests into one model, while allowing the covariates to be misclassified. In particular, we do not require that these parameters be known and provide the means to incorporate information about them from previous studies and expert opinion. We use prior distributions to model our uncertainty about the values of the parameters in both the response model and the measurement error model. In this way, the Bayesian approach provides an attractive method for adjusting inferences to account for measurement error and misclassification. Given that there is no gold standard, our credible intervals do not appear wide, indicating little uncertainty about the prevalence of a disease.

For the simulation example, we chose 0.7 and 0.9 as the true values for specificity and sensitivity, respectively. Typically, psychological measures tend to have

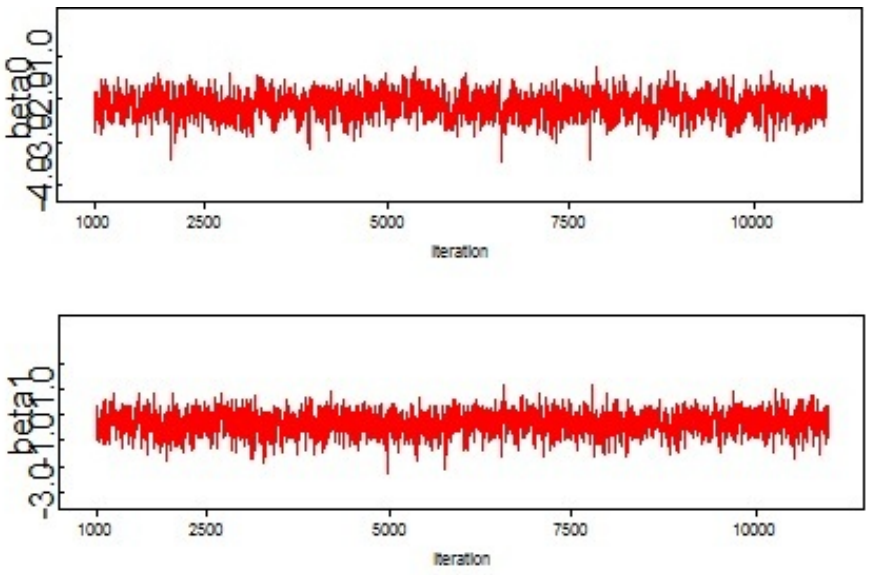

Figure 2.11: Convergence plots for β_0 and β_1 for one data set with highly informative prior variances.

higher specificity because of low base rates, however, we chose to start with lower specificity. This is for comparison purposes.

For the Math example we consider, it appears we need a much larger sample size. Our data set only had 181 subjects, which accounts for the large standard deviations in the results. A larger sample size or more informative priors should yield better results. In Tables 2.8 and 2.9 β_2, β_3 and β_4 are directly comparible, as the represent location of study, high school GPA and gender, respectively. The credible intervals in Table 2.8 for these parameters encompass 0, indicating they are not useful when accounting for misclassification. However, in Table 2.9, when we do not account for misclassification, the credible intervals do not encompass 0 for these parameters. From a (psychological) theoretical perspective gender, overall GPA, and location shouldn't necessarily be good predictors. This may suggest that not accounting for misclassification makes the model susceptible to finding chance-relationship (i.e. Meehl's crud factor) when the variables are measured with error (Meehl (1997)).

There are a number of limitations, indicating the need for further work. First, we assumed that our continuous test results were normally distributed across the population. This assumption is routinely made, but clearly, other distributions can be used (Scott et al. (2008)). We also assumed conditional independence between tests, which may not always hold. If necessary, methods are available that can account for conditional dependence between tests (Dendukuri and Joseph (2001); Black and Craig (2002)), and these can be extended to continuous data. For example, the normal mean can be made a function of the results of one or both of the dichotomous tests. When three conditionally dependent tests are used the model can become non-identifiable, so adding additional tests, if available, could be useful.

2.6 Acknowledgement

The Math Data was collected as part of National Science Foundation Grant 0099216.

CHAPTER THREE

Sample Size Estimation for Independent, Joint Modeling of Efficacy and Safety

Current practice for sample size computations in clinical trials is largely based on frequentist or classical methods. These methods are limited in that they require a point estimate of the variance of the treatment effect. These methods are also based on arbitrary settings of type I and II errors and make no explicit use of prior information (Kikuchi and Gittins (2009)).

The Poisson distribution has a wide spectrum of applications ranging from economics and medicine to actuarial science, and thus, has always been a valuable tool in statistical modeling. One issue of a Poisson setup is the fact that data are often underreported (Stamey and Katsis (2007); Stamey et al. (2004)), such as in environmental and biological data. Failing to account for this underreporting results in biased estimates and inaccurate sample sizes. We illustrate this issue by determining the required sample size to estimate a single Poisson rate utilizing a $(1-\alpha)100\%$ confidence interval. Assuming that no underreporting exists, a frequentist formula yields

$$t = \left(\frac{2z_{\alpha/2}}{\delta}\right)^2 \hat{\lambda} \tag{3.1}$$

where δ denotes the intervals width and $\hat{\lambda}$ is an estimate of λ the unknown rate, $z_{\alpha/2}$ is the upper $(\alpha/2)100\%$ percentile of the standard normal distribution, and t is the necessary sample size. A drawback of this approach is that λ must be estimated in order to determine the sample size. The above equation is modified in the case where not all Poisson events are recorded. Practical situations where this may occur include absences from work (Stamey et al. (2004)) or deaths as recorded on death certificates (Whittemore and Gong (1991)).

In this case, the distribution of $X_1, X_2, \ldots, X_n$ is Poisson(λp). It is straightforward to show that, if p is known, the maximum likelihood estimator for λ is $\hat{\lambda} = \bar{x}/p$, which has a standard error $\sqrt{\lambda/np}$. If p denotes the probability of reporting the event, where $p < 1$ then (3.1) becomes

$$n = \left(\frac{2z_{\alpha/2}}{\delta}\right)^2 \frac{\hat{\lambda}}{p}$$

which results in higher values for n. However, p is rarely known with certainty and the normality assumption for the above equation does not always hold.

Since the reporting probability is usually unknown, the Bayesian approach seems like a natural choice for determining the sample size since it takes into account the uncertainty, which is inherent in any estimation of the unknown parameters. In a Bayesian setting, this uncertainty is expressed through the prior distribution on the parameters of interest.

Stamey et al. (2004) considered a Bayesian analysis for sample size determination as described above. We will extend these methods to the case where we are interested in efficacy and safety. For a drug to be acceptable it must be both safe and effective, and, although some small trade-off between safety and efficacy may be allowable, it is desirable to consider these properties separately when formulating a testing problem (Jennison and Turnbull (1993); Conaway and Petroni (1995), (1996); Bryant and Day (1995)). Jennison and Turnbull (1993) presented sequential designs for bivariate normal endpoints representing treatment efficacy and safety, while Conaway and Petroni (1995) and Bryant and Day (1995) proposed sequential designs for phase II trials that measure antitumor activity and toxicity as binary outcomes

The rest of this chapter is as follows: we describe two models in Section 3.1, one with and one without underreporting. In Section 3.2 we discuss sample size determination and the priors for each parameter in the models. The remaining sections include the results, which includes a Type I error analysis, and a discussion.

3.1 Statistical Methods

Throughout our discussion in this chapter we assume that prior information on the parameters of probability distributions is expressible by means of conjugate prior distributions, which is, frequently, a realistic assumption. One reason for making this assumption is that it simplifies what is, in any case, a complex presentation and ensures that the calculations that we need to make are computationally feasible.

For any sampling distribution, there is a natural family of prior distributions, called the conjugate family. The beta family is conjugate for the binomial family. Thus, if we start with a beta prior, we will end up with a beta posterior. The updating of the prior takes the form of updating its parameters. Mathematically, this is very convenient, for it usually makes calculations quite easy.

For simplicity, we shall also assume that the prior distributions for the difference in efficacy between the new and standard treatments, and for the incidence of adverse events with each of the treatments, are independent. In general, there may be a tendency for more active drugs to be both more efficacious and to cause more adverse reactions, leading to dependent prior distributions. However, there are also many cases for which an assumption of independence is reasonable; published examples include calcineurin inhibitors for immunosuppression in liver transplantation (Perry and Neuberger (2005)), Rosuvastatin to reduce low-density lipoprotein cholesterol (Olsson et al. (2001); Davidson et al. (2002); Saito et al. (2003)), and infliximab (a monoclonal antibody against Tumor Necrosis Factor) for Crohn's disease (Targan et al. (1997)).

The values of the prior distribution hyperparameters, both for efficacy and for the incidence of adverse events, are based on previous experience.

3.1.1 Distribution Theory – Efficacy

Suppose that X and Y are a new treatment and a standard treatment, respectively, and that the treatment groups have the same sample size, $n(>1)$. Let two independent continuous variables X_i and Y_i for $i = 1, 2, \ldots, n$, be the clinical outcomes on some appropriate scale. The subscript i refers to patient i in each treatment group, and X_i and Y_i are unpaired and independent. If $X_i \sim N(\theta+\delta, \sigma^2)$ and $Y_i \sim N(\theta, \sigma^2)$, for $i = 1, 2, \ldots, n$, then,

$$\begin{aligned}
\overline{X} &= \frac{\sum_{i=1}^{i=n} X_i}{n}, \\
\overline{Y} &= \frac{\sum_{i=1}^{i=n} Y_i}{n}, \quad \text{and} \\
\overline{U} &= \overline{X} - \overline{Y}.
\end{aligned}$$

Using the moment generating function, we can find the distribution of $\overline{U}$.

Because we know the distribution of X and Y, we know the moment generating functions of $\overline{X}$ and $\overline{Y}$

$$\begin{aligned}
M_{\overline{X}}(t) &= \exp\left[(\theta+\delta)t + \frac{1}{2n}\sigma^2 t^2\right] \\
M_{\overline{Y}}(t) &= \exp\left[\theta t + \frac{1}{2n}\sigma^2 t^2\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
M_{\overline{U}}(t) &= E\left[\exp\left(\overline{U}t\right)\right] \\
&= E\left[\exp\left[\left(\overline{X}-\overline{Y}\right)t\right]\right] \\
&= \exp\left[(\theta+\delta-\theta)t + \frac{1}{2n}(\sigma^2+\sigma^2)t^2\right] \\
&= \exp\left[\delta t + \frac{1}{2n}2\sigma^2 t^2\right].
\end{aligned}$$

It follows that $\overline{U} \sim N(\delta, 2\sigma^2/n)$. Thus, δ is the mean improvement in efficacy achieved by the new treatment.

Also, writing

$$S_x^2 = \sum_{i=1}^{i=n} \left(X_i - \overline{X}\right)^2,$$
$$S_y^2 = \sum_{i=1}^{i=n} \left(Y_i - \overline{Y}\right)^2, \quad \text{and}$$
$$S^2 = S_x^2 + S_y^2,$$

since $S_x^2/\sigma^2 \sim \chi_{n-1}^2$ and $S_y^2/\sigma^2 \sim \chi_{n-1}^2$ we have $S^2/\sigma^2 \sim \chi_{2n-2}^2$. Let $f(s_x^2, s_y^2, \overline{u_n}|\delta, \sigma^2)$ denote the likelihood function for S_x^2, S_y^2, and $\overline{U_n}$, which is proportional to

$$\left(\sigma^2\right)^{-\frac{(2n-1)}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{s_x^2 + s_y^2 + \frac{n}{2}(\overline{u_n} - \delta)^2\right\}\right]. \tag{3.2}$$

Following O'Hagan (1994), the conjugate prior density function for δ and σ^2 has the form

$$\pi(\delta, \sigma^2) = k(a, g, \omega)\left(\sigma^2\right)^{-\frac{g+3}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{a + (\delta - \mu)^2/\omega\right\}\right]. \tag{3.3}$$

where $k(a, g, \omega) = a^{g/2}2^{-(g+1)/2}(\pi\omega)^{-1/2}\left\{\Gamma(g/2)\right\}^{-1}$, and a, g, and ω are hyperparameters assigned on the basis of prior information. Note that $\pi(\delta, \sigma^2) = \pi(\sigma^2)\pi(\delta|\sigma^2)$, where the distribution of a/σ^2 is chi-squared with g degrees of freedom, and $\pi(\delta|\sigma^2)$ is $N(\mu, \sigma^2\omega)$.

In terms of expectation and variance of the prior distribution for σ^2 we have

$$g = 4 + 2E(\sigma^2)^2/\mathrm{Var}(\sigma^2),$$
$$a = E(\sigma^2)(g-2) \quad \text{and}$$
$$\omega = \tau^2/E(\sigma^2),$$

where τ^2 is the variance of the prior distribution for δ. Applying Bayes theorem and using (3.2) and (3.3), the posterior density for δ and σ^2 may be written as

$$\pi^{(n)}(\delta, \sigma^2|x, y) = k(a', g', \omega')(\sigma^2)^{-(g'+3)/2} \exp\left[-(1/2\sigma^2)\left\{a' + (\delta - \mu')^2/\omega'\right\}\right]. \tag{3.4}$$

Here

$$\omega' = \frac{2\omega}{2+n\omega},$$
$$\mu' = \frac{2\mu + n\omega\overline{z_n}}{2+n\omega},$$
$$g' = g + 2n - 1,$$
$$a' = a + s^2 + \frac{n(\overline{z_n} - \mu)}{2+n\omega}.$$

The mean and variance of the prior distribution for δ are μ and $\omega a/(g-2) = \tau^2$, respectively.

The marginal posterior density function of δ can be obtained by integrating the joint posterior density (3.4) over σ^2 and we have

$$\pi^{(n)}(\delta|x,y) = \frac{1}{B(g'/2,1/2)}(a'\omega')^{-1/2}\left\{1 + \frac{(\delta-\mu')^2}{a'\omega'}\right\}^{-(g'+1)/2}.$$

where

$$1/B(g'/2,1/2) = ([(g'-1)/2]!)/(\sqrt{g'\pi}[(g'-1/2)/2]!).$$

Therefore, $(\delta-\mu')/\sqrt{\omega' z'/g'}$ has a t-distribution with g' degrees of freedom and the mean and variance of the posterior distribution for δ are μ' and $\tau'^2 = \omega' a'/(g'-2)$, respectively.

If the outcomes are binary responses (success of failure) for each patient, with success probabilities P_{X_i} and P_{Y_i} for $i = 1, 2, \ldots, n$, we may convert the outcome to a continuous scale by assuming

$$X_i = \log\frac{P_{X_i}}{1-P_{X_i}} \sim N(\theta+\delta, \sigma^2) \quad \text{and}$$
$$Y_i = \log\frac{P_{Y_i}}{1-P_{Y_i}} \sim N(\theta, \sigma^2) \quad \text{for } i = 1, 2, \ldots, n.$$

3.1.2 Distribution Theory – Adverse Reactions/Safety

Both within the clinical trial and in later use, there are costs associated with adverse events for users of both the current and the new drugs (Kikuchi and Gittins

(2009)). The frequency of adverse reactions also impacts the new drug and whether or not it makes it to market. Assume that adverse reactions occur at the unknown Poisson rate (incidence rate) λ for each patient. We distinguish the new and the current drugs by the subscripts $i = 1$ and 2, respectively. Let r_1 and r_2 denote the number of adverse reactions, for the new and current drugs, respectively, during a total of t patient-years, such that

$$r_i \sim \text{Poisson}(t\lambda_i).$$

Suppose that before the phase II trials λ_i has a prior distribution which is $\Gamma(\alpha_{i0}, \beta_{i0})$, and that during phase II there are r_{i0} adverse events over a total of t_{i0} patient-years. By Bayes theorem it follows that the posterior density for λ_i after phase II is proportional to

$$\begin{aligned}
&\Gamma(\alpha_{i0}, \beta_{i0}) \text{ density} \times P(\text{a Poisson } (\lambda_i t_{i0}) \text{ random variable} = r_{i0}) \\
&\qquad = \frac{\lambda_i^{\alpha_{i0}-1}\beta_{i0}^{\alpha_{i0}}}{(\alpha_{i0}-1)!}\exp\left[-\lambda_i\beta_{i0}\right]\frac{(\lambda_i t_{i0})^{r_{i0}}}{r_{i0}!}\exp\left[-\lambda_i t_{i0}\right] \\
&\qquad \propto \lambda_i^{\alpha_{i0}+r_{i0}-1}\exp\left[-\lambda_i(\beta_{i0}+t_{i0})\right].
\end{aligned}$$

Thus, the posterior distribution for λ_i after phase II is $\Gamma(\alpha_{i0}+r_{i0}, \beta_{i0}+t_{i0})$. This can be used as the prior distribution for λ before phase III, and we write $\alpha_i = \alpha_{i0} + r_{i0}$ and $\beta_i = \beta_{i0} + t_{i0}$. Commonly β_{i0} and α_{i0} are set to small values such as 0.01.

Our posterior distributions for λ_i remain within the gamma family because this is the family of conjugate prior distributions for the parameters of a Poisson process. Note that $\Gamma(\alpha, \beta)$ distribution has mean $\alpha\beta^{-1}$ and variance $\alpha\beta^{-2}$.

3.1.3 Model 1 – Without Underreporting

In this chapter, we will look at two models, one in which we take into account that adverse events are often underreported and one in which we do not. If we do not account for underreporting, the model is as described above.

3.1.4 Model 2 – With Underreporting

In Model 2, we will take into account that adverse reactions are often underreported. Failing to account for underreporting may result in inaccurate estimates. Inaccurate estimates may, in turn, be costly in an experiment. The Poisson model we consider here is the following

$$w_i \sim \text{Poisson}(t\lambda_i p_i), \quad i = 1, 2, \tag{3.5}$$

where w_i represents the underreported counts. Now, while the true number of occurrences is r, we only observe w of these.

Here, t is the sample size, sometimes referred to as the opportunity size because it is often an area or length of time, λ_i is the Poisson rate of the ith population and p_i is the probability that a particular occurrence is observed in the ith sample, referred to as the reporting probability. Assuming the reporting probability is the same in both populations reduces the amount of variability in the estimators and generally would lead to a smaller required sample size, but this could be a strong assumption.

This model is an extension of Stamey et al. (2004), who determined the required fallible sample size for the one sample case. We use the same prior structure as in Fader and Hardie (2000), placing a beta distribution on p and a gamma distribution on λ:

$$\pi_p(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \quad a, b > 0$$
$$\pi_\lambda(\lambda) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-\lambda/\beta}, \quad \alpha, \beta > 0.$$

The derived posterior distributions are not in a true "closed form" since they are functions of hypergeometric functions and confluent hypergeometric functions, which are infinite sums. Specifically, the posterior for λ is

$$\pi(\lambda_i|w_i) = \frac{(t_i+\beta_i)^{w_i+\alpha_i}\lambda_i^{w_i+\alpha_i-1}}{\Gamma(w_i+\alpha_i)} \frac{{}_1F_1(w_i+a_i, w_i+a_i+b_i, -t_i\lambda_i)}{{}_2F_1(w_i+\alpha_i, b_i, w_i+a_i+b_i, t_i/(t_i+\beta_i))}$$

where ${}_1F_1$ is the confluent hypergeometric function, ${}_2F_1$ is the Gauss hypergeometric function, and $t_1 = t_2$ (Anderson et al. (1994)). These forms are not particularly useful in a simulation-based sample size determination procedure; thus, we use the Gibbs sampler to estimate the posterior densities.

The Gibbs sampler is a special case of Metropolis-Hastings sampling. The Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. This is particularly useful when direct sampling is difficult. Through the use of techniques like the Gibbs sampler, we can avoid difficult calculations, replacing them instead with a sequence of easier calculations.

For the Gibbs sampler, we augment the observable data with the latent variables Z_i which are the unobserved underreported number of occurences in population i. Combining the Poisson data in (3.5) with the above conjugate priors and the latent data yields the following joint posterior:

$$\pi(\lambda_1, \lambda_2, p_1, p_2, Z_1, Z_2 | w_1, w_2) \propto \prod_{i=1}^{2} p_i^{w_i + a_i - 1} (1 - p_i)^{Z_i + b_1 - 1} e^{-(t_i + \beta_i)\lambda_i} \lambda_i^{w_i + Z_i + \alpha_i - 1}. \tag{3.6}$$

The full likelihood contains the x's and y's, and because they are independent, the full likelihood is just the product of the two components.

If the reporting probabilities are assumed to be the same, the joint posterior simplifies slightly since there is one less unknown parameter. To implement the Gibbs sampler, we rearrange and manipulate (3.6), which yields

$$\begin{aligned} \lambda_i | Z_i, p_i &= c_1 \lambda_i^{w_i + Z_i + \alpha_i - 1} e^{-(t_i + \beta_i)\lambda_i}, \\ p_i | Z_i, \lambda_i &= c_2 p^{w_i + a_i - 1} (1 - p)^{Z_i + b_i - 1}, \quad \text{and} \\ Z_i | \nu, \tau, \lambda, p, \mathbf{x} &= c_3 \frac{\left[(1 - p)\, \lambda_i \right]_i^Z}{Z_i!} \end{aligned}$$

where

$$c_1 = \frac{t_i + \beta_i}{\Gamma(w_i + Z_i + \alpha_i)},$$
$$c_2 = \frac{\Gamma(w_i + a_i + Z_i + b_i)}{\Gamma(w_i + a_i)\Gamma(Z_i + b_i)}, \quad \text{and}$$
$$c_3 = e^{-(1-p_i)\lambda_i}.$$

Thus, the following full conditionals are required:

$$\lambda_i | Z_i, p_i \sim \text{Gamma}(w_i + Z_i + \alpha_i, t_i + \beta_i)$$
$$P_i | Z_i, \lambda_i \sim \text{Beta}(w_i + a_i, Z_i + b_i)$$
$$Z_i | \nu, \tau, \lambda, p, \mathbf{x} \sim \text{Poisson}(t_i \lambda_i (1 - p_i))$$

where $t_1 = t_2$.

After a suitable burn-in, sampling iteratively from the above distributions yields a Markov Chain Monte Carlo (MCMC) approximation to the posterior distribution. From this chain, quantities such as the ratio, λ_1/λ_2, or the difference, $\lambda_1 - \lambda_2$, may be approximated as well.

3.2 Sample Size Determination

In this section we overview the simulation-based procedure of Wang and Gelfand (2002) and apply it to our model. This model was also recently applied by Beavers and Stamey (2012). Throughout this section we assume interest lies in the posterior distribution of the difference of the two normal means, δ_μ, for the efficacy variable and the difference of the two rates, $\lambda_1 - \lambda_2$, for the safety variable.

For frequentist sample size approaches, unknown parameters and effect sizes are replaced with fixed estimates that are elicited from experts, based on pilot data, or chosen for their conservative performance. Instead of plugging in fixed numbers for inputs, Wang and Gelfand (2002) suggest eliciting probability distributions that allow for uncertainty in these estimates. These distributions are referred to as design

priors. The design priors are usually required to be at least moderately informative, representing the assumed true nature of the data, as opposed to the prior distributions used for data analysis, or analysis priors, which can be either informative or diffuse.

We assume that a specific effect for $\lambda_1 - \lambda_2$ is available for our sample size determination procedure. For instance, we believe there is a difference in the rates of 2, so we fix this value in the design phase. Likewise, for δ_μ, we believe there is a difference in the means of 5, so we fix this value in the design phase. For the rest of the parameters, we recommend using design priors that elicit "most likely" values according to prior knowledge, along with a range in which these parameters are most likely to fall.

3.2.1 Type I Error Analysis

Interest lies in determining the required sample size to show that a parameter is "significantly" different from 0. A threshold other than 0 would be straightforward to incorporate into the procedure. If a positive relationship between the difference of the normally distributed responses is expected, a Bayesian power criterion selects n so that, for probabilities α and η,

$$E\left[I\left\{Pr\left(\delta_\mu > 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta \tag{3.7}$$

where $I\{\}$, is the indicator function, $\delta_\mu = \mu_X - \theta$, and $\mathbf{d^{(n)}}$ represents a generated data set of size n. Common choices for α are 0.1, 0.05, and 0.01 while η is typically 0.8 or 0.9. Here, the expectation is with respect to the design prior and the posterior probability, which is the argument of the indicator function. For each data set, $\mathbf{d^{(n)}}$, the null hypothesis of $H_0 : \delta_\mu \leq 0$ is rejected in favor of $H_1 : \delta_\mu > 0$ if $Pr(\delta_\mu > 0|\mathbf{d^{(n)}}) > 1-\alpha$. With equation (3.7), we seek the sample size for which such hypotheses are rejected at least $100(1-\eta)\%$ of the time. If the relationship

with the covariate is expected to be negative, equation (3.7) becomes

$$E\left[I\left\{Pr\left(\delta_\mu < 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta.$$

Results are displayed in Figure 3.1(left).

If interest is in the treatment/safety relationship, a similar criterion is used. In the same degree, if a positive relationship between the difference of the normally distributed responses is expected, a Bayesian power criterion selects t so that, for probabilities α and η,

$$E\left[I\left\{Pr\left(\delta_\lambda > 0|\mathbf{d^{(t)}}\right) > 1-\alpha\right\}\right] \geq \eta \tag{3.8}$$

where $\delta_\lambda = \lambda_1 - \lambda_2$ and $\mathbf{d^{(t)}}$ represents a generated data set of size t. Figure 3.1(right) displays these results.

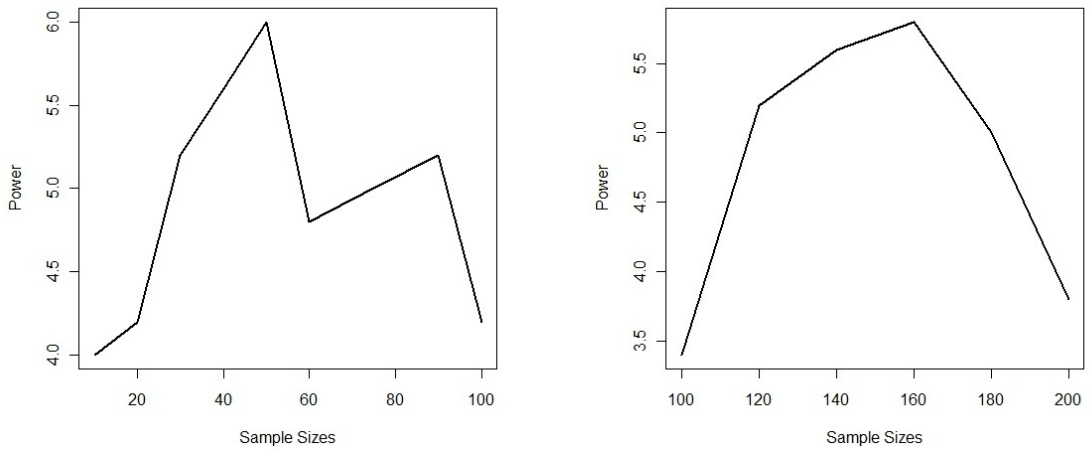

Figure 3.1: Type I error analysis results for δ_μ when $\delta_\mu = 0$ (left) and for δ_λ when $\delta_\lambda = 0$ for the model not accounting for underreporting.

These procedures can be extended to the situation of testing multiple hypotheses simultaneously. For instance, testing both $H_0 : \delta_\mu = 0$ and $H_0 : \delta_\lambda = 0$ may be of interest. Extension can be made to other multiple hypothesis structures as well. The goals of multiple testing need to be carefully considered. For instance, Sozu et al. (2011) provides a sample size formula for testing superiority with multiple end-points. In their case, the overall null hypothesis is rejected if each of the

individual components is rejected and sample size is determined to achieve certain power to reject the overall null hypothesis. In our case, this would correspond to

$$E\left[I\left\{Pr\left(\delta_\mu > 0|\mathbf{d^{(n)}}\right) > 1-\alpha \cap Pr\left(\delta_\lambda > 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta. \tag{3.9}$$

This has a large impact on Type II error, but Type I error will be quite small using this framework. Figure 3.2 displays these results. An alternative would be to find the sample sizes for both criteria (3.7) and (3.8) and choose the larger. This approach tends to inflate Type I error, so the value of α may have to be adjusted slightly. For instance, the value of α may be verified by checking via simulation to assure that it is controlled at a reasonable level. These results are in Figure 3.2.

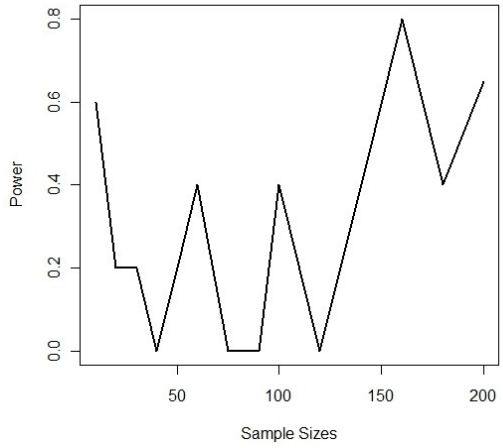

Figure 3.2: Type I error analysis results for the model not accounting for underreporting when $\delta_\mu = 0$ and $\delta_\lambda = 0$.

The computing algorithm for Bayesian power is given below, and we wish to determine the total sample size n that satisfies our power criteria. The expected power is approximated for each sample size based on B simulated data sets. In the following sequence, the subscript $k \in \{1, \ldots, B\}$ refers to the iteration in the simulation for each sample size. The following steps are used for a single value of n, and they are repeated across a grid of potential sample sizes.

(1) Generate a single value for each parameter value λ_1, λ_2, θ, and μ_X from appropriate distributions.

(2) For sample size n, simulate error-free data X_i, Y_i, and r_i, $i = 1, \ldots, n$ from a suitable distribution over its expected range.

(3) Fit the Bayesian model to the simulated data generated in steps 1 and 2 using the analysis priors and approximate the posterior distribution of $\lambda_1 - \lambda_2$, including the alternative posterior probability $Pr(\lambda_1 - \lambda_2 \neq 2|\mathbf{d^{(n,k)}})$, where $\mathbf{d^{(n,k)}})$ denotes the data generated at the kth iteration for sample size n.

(4) Repeat steps 1 - 3 B times at each sample size value n, each time storing $Pr(\lambda_1 - \lambda_2 \neq 2|\mathbf{d^{(n,k)}})$.

(5) Calculate the posterior probability that $Pr(\lambda_1 - \lambda_2 \neq 2|\mathbf{d^{(n,k)}}) > 1 - \alpha$ via the formula

$$m^{(n)} = \frac{1}{B} \sum_{k=1}^{B} I\{Pr(\lambda_1 - \lambda_2 \neq 2|\mathbf{d^{(n,k)}}) > 1 - \alpha\}.$$

(6) Finally, repeat steps 1 - 5 for a range of sample sizes and plot $m^{(n)}$ by n to find a sample size that achieves a desired level of power.

The same process was applied to the normal differencing, $\mu_X - \theta$. Therefore, the probability of interest in step 3 becomes $Pr(\mu_X - \theta \neq 5|\mathbf{d^{(n,k)}})$.

The method was performed using the software packages R and WinBUGS. These packages are freely available on the internet, and the code is available in Appendix B.

3.2.2 Prior Distributions

To complete the Bayesian model, we require prior distributions for the model parameters. The sample size determination approach we apply here is similar to that proposed by Wang and Gelfand (2002) in which multiple data sets are simulated and

then fit using a Bayesian model. The simulation-based approach requires two sets of prior distributions. One set, known as the design priors, is discussed in the next section. The other set, the anlaysis priors, are the priors used in the data anlaysis of the simulation-based sample size determination scheme and would be used to analyze the data when the study is actually performed.

In the absence of relevant prior data or expert opinion, diffuse normal prior distributions are often employed as analysis priors for the means of the data.

3.2.2.1 *Model 1 – Without Underreporting.* For the first model, where we do not account for underreporting in the safety variable, we determine the required sample size to obtain a $(1-\alpha)100\%$ posterior interval for δ_μ and δ_λ. For efficacy, the data are described in Section 3.1.1, where X and Y are Normal, and $\delta_\mu = \mu_X - \theta$, the difference in the Normal means. The parameters above will have the following diffuse priors:

$$\theta \sim \text{Normal}(0,\ 0.0001)$$

$$\mu_X \sim \text{Normal}(0,\ 0.0001)$$

$$\sigma^2 \sim \text{Uniform}(0.1,\ 50).$$

For the safety variable, the data are described in Section 3.1.2, where r_i represents the number of adverse events and $\delta_\lambda = \lambda_1 - \lambda_2$. The parameters have the following priors:

$$\lambda_i \sim \text{Gamma}(0.01,\ 0.01)$$

$$\delta_\lambda = \lambda_1 - \lambda_2.$$

Prior distributions are based on previous experience or expert opinion. We are interested in δ_μ and δ_λ, the mean improvement in efficacy and safety, respectively, as well as the power associated with each. A curve can then be fit using the sample sizes and corresponding powers.

3.2.2.2 *Model 2 – With Underreporting.* For the case in which we do not account for underreporting, we determine the required sample size to obtain a $(1-\alpha)100\%$ posterior interval for δ_μ and δ_λ. For efficacy, the data are defined in Section 3.1.1, and the priors are the same as the model without underreporting (described in Section 3.2.1.1). For the safety variable, the data are defined in Section 3.1.4, where w_i are the observed number of underreported adverse events. The priors are

$$p \sim \text{Beta}(50,10)$$

$$\lambda_i \sim \text{Gamma}(0.01,\ 0.01)$$

$$\delta_\lambda = \lambda_1 - \lambda_2,$$

where p is the same for the new and standard drug. We allow r_i to be the true number of adverse reactions, which we do not know in this case. Again, prior distributions are based on previous experience or expert opinion. We are interested in δ_μ and δ_λ, the mean improvement in efficacy and safety, respectively, as well as the power associated with each. A curve can then be fit using the sample sizes and corresponding powers to find the total required sample size.

3.2.3 MCMC Implementation

In the analysis of our data, we utilized a flexible software for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods. This iterative algorithm builds approximations to the posterior distributions of interest based on Monte Carlo simulations. Compared with alternatives such as numerical integration, MCMC methods require much less computing times and are easier to implement and customize, especially using the package WinBUGS.

All simulations in Section 3.3 used $B = 500$ datasets generated at each sample size, with sample sizes ranging from 5 to 200 in varying increments. We produce posterior approximations and probabilities based on MCMC samples of 10000 posterior iterates after a 1000 iteration burn-in for each generated dataset.

3.3.1 Model 1 – Without Underreporting

Using the data and priors described in Section 3.2.1, we obtain the following power curves for δ_μ, the difference in Normal means, and δ_λ, the difference in Poisson rates. In Figure 3.3(left), we note that with the efficacy power curve, roughly 150 participants are needed for a power of 80%. However, the safety curve, Figure 3.3(right), shows a necessary sample size of 15 person-years (or some other appropriate person-time) to reach a power of 80%.

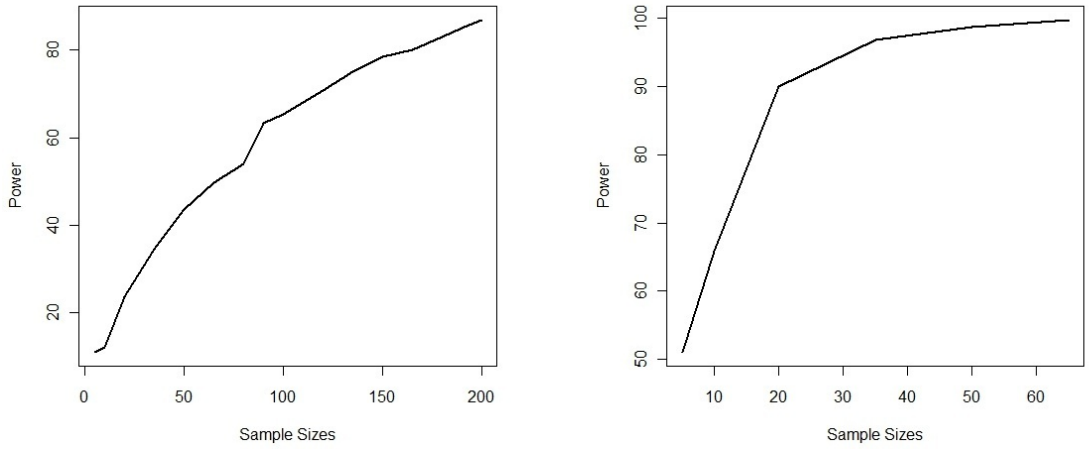

Figure 3.3: Simulation results for δ_μ (left) and δ_λ (right) for the model not accounting for underreporting.

3.3.2 Model 2 – With Underreporting

Using the data and priors described in Section 3.2.2, we obtain the following power curves for δ_μ, the difference in Normal means, and δ_λ, the difference in Poisson rates. In Model 2, we arbitrarily chose parameters for p, the reporting probability, to be roughly 85%. In other words, we expect 85% of adverse events to be reported. For efficacy, in Figure 3.4(left), roughly 150 participants are necessary for a power of 80%, as with the previous model. However, for the safety variable the necessary sample size is now 20 person-years for a power of 80% in Figure 3.4(right).

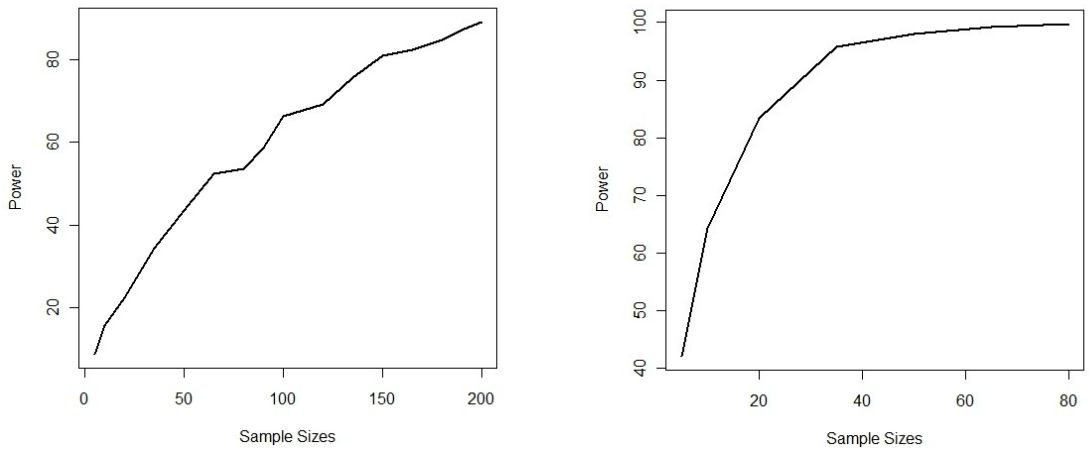

Figure 3.4: Simulation results for δ_μ (left) and δ_λ (right) for the model accounting for underreporting.

Our results did not change for efficacy because we did not alter our model for efficacy. The data was constructed from the same distributions, and the priors did not change. We expected the results for efficacy to be very similar from Model 1 to Model 2. The safety results, on the other hand, did change. For the same power, our sample size increased from Model 1 (no underreporting) to Model 2 (with underreporting).

Recall that p is the reporting probability, and we assume that p is the same for both treatment groups in Model 2. The reporting probability can have a great effect on the desired outcome. A higher reporting probability represents more reported adverse reactions. A higher reporting of adverse events would lead to more accurate sample sizes. Conversely, a lower reporting probability represents fewer reported adverse events and adds to the posterior variability. Table 3.1 below shows the power and sample sizes associated with different reporting probability parameters. Across the top are different priors for the reporting probability, and down the left side are different sample sizes.

We can see that as the reporting probability parameters increase, so does the power associated with each opportunity size. We expect this phenomenon to be

Table 3.1: Simulation Results For δ_λ for the Model Not Accounting for Underreporting With Different Reporting Probability Parameters.

t	Beta(10,10)	Beta(30,10)	Beta(50,10)	Beta(70,10)	Beta(90,10)
5	31.4	39.0	39.4	38.2	43.4
10	45.8	61.6	62.6	64.2	61.8
20	64.4	82.2	84.6	85.0	82.8
35	81.8	91.4	96.8	95.6	96.6
50	90.6	98.8	99.0	99.4	98.0

the case because a higher parameter value indicates a greater probability of adverse events being reported. For instance, if the prior for p is a Beta(90,10), we can expect an average of $90/(90 + 10) = 90\%$ of adverse events to be reported. On the other hand, if the prior for p is a Beta(10,10), we expect only an average of $10/(10 + 10) = 50\%$ of adverse events to be reported. If all of the adverse events are reported, our results are similar to those of Model 1, in which we do not account for underreporting. However, since we do not always have complete reporting of adverse reactions, our results vary slightly.

We also note that as our opportunity size increases within each reporting probability prior, our power increases as well. For instance, the column with prior parameters 50 and 10 indicate a mean reporting of adverse events to be $50/(50+10) = 83\%$. The power greatly increases as the opportunity size increases from 5 to 50.

3.4 Discussion

In this chapter, our models accurately estimate the difference in the means of the data as well as the difference in the Poisson rates, while allowing efficacy and safety to be independent. The model with underreporting was found to increase the needed sample size.

In general, a tendency may exist for more active drugs to be both more efficacious and to cause more adverse reactions, leading to dependent prior distributions.

However, many cases also exist for which an assumption of independence is reasonable. In the next chapter, we will consider a model where efficacy and safety are dependent. This dependence will be depicted using regression modeling. We will, again, account for underreporting of adverse events, and we will compare our results to a model not accounting for underreporting. We expect that this underreporting will affect the power curve for efficacy as well as for safety.

Due to the computational requirements, Monte Carlo or some other method of posterior approximation is required to estimate the sample size. We have considered the effect of prior information about p on the required sample size. Gains are more pronounced as the reporting probability nears 1. Finally, note that in some applications, the sample sizes are in terms of linear feet and time, not a number of observations. For these applications the sample would not need to be constrained to an integer.

CHAPTER FOUR

Sample Size Estimation for Dependent, Joint Modeling of Efficacy and Safety

Poisson data with underreporting is a well-researched problem due to its applicability in many fields, including economics, epidemiology, and actuarial science. The increased uncertainty due to the underreporting causes traditional methods of sample size determination to underestimate the needed sample size.

In this chapter, we will look at a problem similar to that of Kikuchi and Gittins (2009) using different sample size methods. Kikuchi and Gittins (2009) used a behavioral Bayes method to determine the sample size of a clinical trial, taking into effect efficacy and safety. Their paper followed that of Gittins and Pezeshk (2000), which introduced a fully Bayesian approach to sample size determination in clinical trials. In contrast to the usual Bayesian decision theoretic methodology, which assumes a single decision maker, the Gittins and Pezeshk approach recognizes the existence of three decision makers, namely: the pharmaceutical company conducting the trial, which decides on its size; the regulator, whose approval is necessary for the drug to be licensed for sale; and the public at large, who determine ultimate usage. Moreover, they model the subsequent usage by plausible assumptions for actual behaviour, rather than assuming that it represents decisions which are in some sense optimal. Their results show that the optimal sample size depends strongly on the expected benefit from a conclusively favourable outcome, and on the strength of the evidence required by the regulator.

In this chapter, we will use methods similar to that of Chapter 3, but we will allow efficacy and safety to be dependent. This is an important extension. We chose to apply this dependence using regression modelling. However, we could use a random effects model to account for the dependence (Hedeker et al. (1994)).

The rest of this chapter is as follows: we describe two models in Section 4.1, one with and one without underreporting. In Section 4.2 we discuss sample size determination and the priors for each parameter in the models. The results are given in Section 4.3, which includes analysis of different prior parameters as well as a Type I error analysis. The final section contains a discussion.

4.1 Statistical Methods

We assume that the difference in efficacy between the new and standard treatments, and for the incidence of adverse reactions with each of the treatments, are dependent. In general, there may be a tendency for more active drugs to be both more efficacious and to cause more adverse reactions, leading to dependent prior distributions. This dependence will be expressed through a regression model.

The values of the prior distribution hyperparameters, both for efficacy and for the incidence of adverse reactions, are based on previous experience.

4.1.1 Distribution Theory – Efficacy

Suppose that

$$Y_i \sim \text{Normal}(\theta_i, \sigma^2)$$

for $i = 1, 2, \ldots$, be the clinical outcomes on some appropriate scale, such that

$$\theta_i = \beta_0 + \beta_1 Z(i),$$

where Z is an indicator such that if $Z = 0$, the data is from the standard treatment, and if $Z = 1$, the data is from the new treatment. The subscript i refers to patient i in each treatment group. It would be straightforward to include more covariates, $\theta_i = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$.

4.1.2 Distribution Theory – Adverse Reactions/Safety

Both within the clinical trial and in later use, there are costs associated with adverse events for users of both the current and the new drugs (Kikuchi and Gittins

(2009)). Their frequency also impacts the new drug and whether it makes it to market. Assume that adverse reactions occur at the unknown Poisson rate (incidence rate) λ for each patient. We distinguish the new and the current drugs by the subscripts $i = 1$ and 2, respectively.

Let

$$r_i \sim \text{Poisson}(\mu_{r_i})$$

denote the numbers of adverse events, for the new and current drugs during a total of t patient-years, such that

$$\log(\mu_{r_i}) = \gamma_0 + \gamma_1 Z(i) + \gamma_2(Y_i - \theta),$$

where γ_2 allows for correlation between the efficacy and safety variables.

Note our rate is an exponential function. Thus, the difference in the treatment groups is now $\exp(\gamma_1)$, and the treatment group has an increase in the rate of adverse events of $\exp(\gamma_1)$. For instance, if $\gamma_1 = 0.6$, our rate of increase is $\exp(0.6) = 1.82$. Likewise, if we are interested in how large the sample needs to be for an increase in rate of 2, then $\gamma_1 = \log(2) = 0.693$.

4.1.3 Model 1 – Without Underreporting

In this chapter, we will look at two models, one in which we take into account that adverse events are often underreported and one in which we do not. If we do not account for underreporting, the model is as described above. We will use Poisson priors for the safey variable.

4.1.4 Model 2 – With Underreporting

In Model 2, we will take into account that adverse events are often underreported. Failing to account for underreporting may result in inaccurate estimates. Inaccurate estimates may, in turn, be costly in an experiment. The Poisson model

we consider here is the following

$$w_i \sim \text{Poisson}(\mu_w), \quad i = 1, 2, \tag{4.1}$$

where w_i represents the underreported counts. We define the Poisson rate as:

$$\mu_w = p\mu_{r_i}, \quad \text{where}$$
$$\log(\mu_{r_i}) = \gamma_0 + \gamma_1 Z_i + \gamma_2(Y - \mu_Y).$$

Here, p_i is the probability a particular occurrence is observed in the ith population, referred to as the reporting probability, γ_i are the coefficients, Z_i is the indicator for the data, and Y are the data for efficacy. Assuming the reporting probability is the same in both populations reduces the amount of variability in the estimators and generally would lead to a smaller required sample size, but this could be a strong assumption.

This model is an extension of Chapter Three, in which we determine the necessary sample size for independent efficacy and safety. In this chapter, we allow efficacy and safety to be dependent on each other. We use the same prior structure as in Fader and Hardie (2000), placing a beta distribution on p and a gamma distribution on λ:

$$\pi_p(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \quad a, b > 0$$
$$\pi_\lambda(\lambda) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-\lambda/\beta}, \quad \alpha, \beta > 0.$$

The derived posterior distributions are not in a true "closed form" since they are functions of hypergeometric functions and confluent hypergeometric functions, which are infinite sums. Specifically, the posterior for λ is

$$\pi(\lambda_i|w_i) = \frac{(t_i+\beta_i)^{w_i+\alpha_i}\lambda_i^{w_i+\alpha_i-1}}{\Gamma(w_i+\alpha_i)} \frac{{}_1F_1(w_i+a_i, w_i+a_i+b_i, -t_i\lambda_i)}{{}_2F_1(w_i+\alpha_i, b_i, w_i+a_i+b_i, t_i/(t_i+\beta_i))}$$

where ${}_1F_1$ is the confluent hypergeometric function, ${}_2F_1$ is the Gauss hypergeometric function, and $t_1 = t_2$. These forms are not particularly useful in a simulation-based

sample size determination procedure; thus, we use the Gibbs sampler to estimate the posterior densities.

The Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. This is particularly useful when direct sampling is difficult. Through the use of techniques like the Gibbs sampler, we can avoid difficult calculations, replacing them instead with a sequence of easier calculations. As with other MCMC algorithms, Gibbs sampling generates a Markov chain of samples.

For the Gibbs sampler, we augment the observable data with the latent variables Z_i which are the unobserved underreported number of occurences in population i. Combining the Poisson data in (4.1) with the above conjugate priors and the latent data yields the following joint posterior:

$$\pi(\lambda_1, \lambda_2, p_1, p_2, Z_1, Z_2 | w_1, w_2) \propto \prod_{i=1}^{2} p_i^{w_i + a_i - 1} (1 - p_i)^{Z_i + b_1 - 1} e^{-(t_i + \beta_i)\lambda_i} \lambda_i^{w_i + Z_i + \alpha_i - 1}. \quad (4.2)$$

After a suitable burn-in, sampling iteratively from the above distributions yields an Markov Chain Monte Carlo (MCMC) approximation to the posterior distribution. From this chain, quantities for the coefficients β_1 and γ_1 may be approximated.

4.2 Sample Size Determination

In this section we overview the simulation based procedure of Beavers and Stamey (2012) and apply it to our model. Throughout this section we assume interest lies in the posterior distribution of the difference of the two rates, β_1.

For frequentist sample size approaches, unknown parameters and effect sizes are replaced with fixed estimates that are elicited from experts, based on pilot data, or chosen for their conservative performance. Instead of plugging in fixed numbers for inputs, Wang and Gelfand (2002) suggest eliciting probability distributions that

allow for uncertainty in these estimates. These distributions are referred to as design priors. The design priors are usually required to be at least moderately informative, representing the assumed true nature of the data, as opposed to the prior distributions used for data analysis, or analysis priors, which can be either informative or diffuse.

We assume that a specific effect for β_1 is available for our sample size determination procedure. For instance, we believe the treatment effect has an increase of 3 units, so we fix this value in the design phase. For the rest of the parameters, we recommend using design priors that elicit "most likely" values according to prior knowledge, along with a range in which these parameters are most likely to fall.

The computing algorithm for Bayesian power is given below, and we wish to determine the total sample size n that satisfies our power criteria. The expected power is approximated for each sample size based on B simulated data sets. In the following sequence, the subscript $k \in \{1, \ldots, B\}$ refers to the iteration in the simulation for each sample size. The following steps are used for a single value of n, and they are repeated across a grid of potential sample sizes.

(1) Generate a single value for each covariate β_0, β_1, γ_0, γ_1 and γ_2.

(2) For sample size n, simulate error-free data X_i, Y_i, and w_i, $i = 1, \ldots, n$ from a suitable distribution over its expected range.

(3) Fit the Bayesian model to the simulated data generated in steps 1 and 2 using the analysis priors and approximate the posterior distribution of β_1, including the alternative posterior probability $Pr(\beta_1 > 0|\mathbf{d^{(n,k)}})$, where $\mathbf{d^{(n,k)}})$ denotes the data generated at the kth iteration for sample size n.

(4) Repeat steps 1 - 3 B times at each sample size value n, each time storing $Pr(\beta_1 > 0|\mathbf{d^{(n,k)}})$.

(5) Calculate the posterior probability that $Pr(\beta_1 > 0|\mathbf{d^{(n,k)}}) > 1 - \alpha$ via the formula

$$m^{(n)} = \frac{1}{B}\sum_{k=1}^{B} I\{Pr(\beta_1 > 0|\mathbf{d^{(n,k)}}) > 1 - \alpha\}.$$

(6) Finally, repeat steps 1 - 5 for a range of sample sizes and plot $m^{(n)}$ by n to find a sample size that achieves a desired level of power.

The same process was applied to the normal differencing, $\mu_X - \theta$. Therefore, the probability of interest in step 3 becomes $Pr(\mu_X - \theta \neq 5|\mathbf{d^{(n,k)}})$.

The method was performed using the software packages R and WinBUGS. These packages are freely available on the internet, and the code is available in Appendix C.

4.2.1 Prior Distributions

To complete the Bayesian model, we require prior distributions for the model parameters. The sample size determination approach we apply here is similar to that used in Chapter Three, which was based on the methds proposed by Wang and Gelfand (2002). These methods involve simulating multiple data sets and then fit them using a Bayesian model. The simulation-based approach requires two sets of prior distributions. One set, known as the design priors, is discussed in the next section. The other set, the anlaysis priors, are the priors used in the data anlaysis of the simulation-based sample size determination scheme and would be used to analyze the data when the study is actually performed.

In the absence of relevant prior data or expert opinion, diffuse normal prior distributions are often employed as analysis priors for the means of the data.

4.2.1.1 *Model 1 – Without Underreporting.* For the first model in which we do not account for underreporting in the safety variable, we determine the required sample size to obtain a $(1-\alpha)100\%$ posterior interval for β_1 and γ_1. For efficacy, the

data are described in Section 4.1.1. The parameters will have the following diffuse priors:

$$\beta_0 \sim \text{Normal}(0,0.01)$$

$$\beta_1 \sim \text{Normal}(0,0.01)$$

$$\sigma^2 \sim \text{Uniform}(0.1,\ 50).$$

For safety, the data are described in Section 4.1.2, where r_i represents the number of adverse events, γ_i are the coefficients, Z_i is the data indicator, and Y are the data. The parameters will have the following diffuse priors:

$$\gamma_0 \sim \text{Normal}(0,0.1)$$

$$\gamma_1 \sim \text{Normal}(0,0.1)$$

$$\gamma_2 \sim \text{Normal}(0,0.1).$$

We are interested in β_1 and γ_1, the mean improvement in efficacy and safety, respectively, as well as the power associated with each. A curve can then be fit using the sample sizes and corresponding power.

4.2.1.2 *Model 2 – With Underreporting.* For the case of not accounting for underreporting we determine the required sample size to obtain a $(1-\alpha)100\%$ posterior interval for β_1 and γ_1. For efficacy, the data are, again, described in Section 4.1.1, and the priors are discussed in Section 4.2.1.1. For safety, the data are described in Section 4.1.4, where w_i are the observed number of underreported adverse reactions. The parameters will have the following diffuse priors:

$$\gamma_0 \sim \text{Normal}(0,0.1) \qquad \gamma_1 \sim \text{Normal}(0,0.1)$$

$$\gamma_2 \sim \text{Normal}(0,0.1) \qquad p \sim \text{Beta}(50,\ 10).$$

and p is the same for the new and standard drug. We allow r_i to be the true number of adverse reactions, which we do not know in this case. Again, prior distributions

are based on previous experiments or expert opinion, and we are interested in β_1 and γ_1, the mean improvement in efficacy and safety, respectively, as well as the power associated with each. A curve can then be fit using the sample sizes and corresponding power.

4.2.2 MCMC Implementation

In the analysis of our data, we utilized a flexible software for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods. This iterative algorithm builds approximations to the posterior distributions of interest based on Monte Carlo simulations. Compared with alternatives such as numerical integration, MCMC requires much less computing times and is easier to implement and customize.

All simulations in the next section used $B = 500$ datasets generated at each sample size from 5 to 200 in varying increments. We produce posterior approximations and probabilities based on Markov Chain Monte Carlo samples of 10000 posterior iterates after a 1000 iteration burn-in for each generated dataset.

4.3 Results

4.3.1 Model 1 – Without Underreporting

Using the data and priors described in Section 4.2.1, we obtain the following power curves for β_1, the coefficient for the mean of the Normal data, and γ_1, the coefficient for the mean of the Poisson rate. We note that for the efficacy variable, in Figure 4.1(left), 175 participants are needed for a power of 80%. However, the safety power curve, in Figure 4.1(right), shows a necessary sample size of 75 person-years (or some other appropriate person-time) to reach a power of 80%.

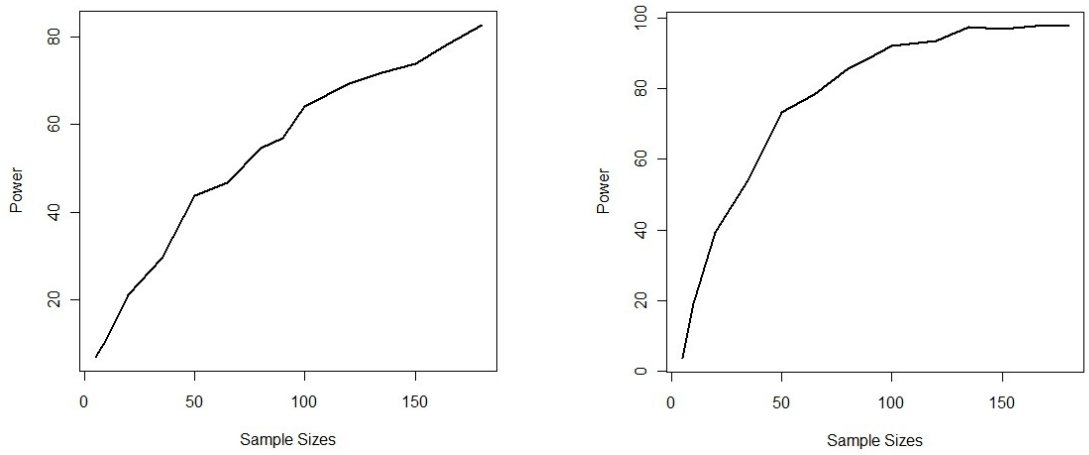

Figure 4.1: Simulation results for β_1 (left) and γ_1 (right) for the model not accounting for underreporting.

4.3.2 Model 2 – With Underreporting

Using the data and priors described in Section 4.2.2, we obtain the following power curves for β_1, the coefficient for the mean of the Normal data, and γ_1, the coefficient for the mean of the Poisson rate. We note that in Figure 4.2(left), for the efficacy variable, 200 participants are needed for a power of 70%. However, the safety power curve shows a necessary sample size of 100 person-years (or some other appropriate person-time) to reach a power of 80% in Figure 4.2(right).

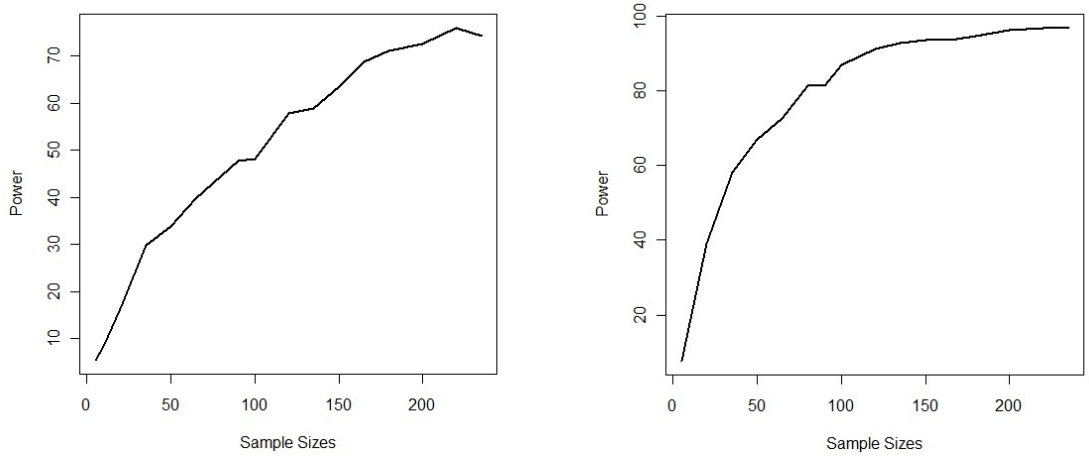

Figure 4.2: Simulation results for β_1 (left) and γ_1 (right) for the model accounting for underreporting.

We note that in Chapter Three our efficacy sample size did not change when underreporting was accounted for in the model. In this chapter, our efficacy sample size did change. Now that efficacy and safety are dependent through a regression model, as one changes, so does the other. In Model 2, we need more participants to reach a smaller power for efficacy. For safety, we also need a larger sample size.

Recall that p is the reporting probability. We assume that p is the same for both treatment groups in Model 2. The reporting probability can have a great effect on the desired outcome. A higher reporting probability represents more reported adverse events. A higher reporting of adverse events would lead to more accurate sample sizes. Table 4.1 below shows the power and sample sizes associated with different reporting probability parameters.

Table 4.1: Simulation Results for γ_1 for the Model Not Accounting for Underreporting With Different Reporting Probability Parameters.

n	Beta(10,10)	Beta(30,10)	Beta(50,10)	Beta(70,10)	Beta(90,10)
50	55.8	63.6	69.2	69.6	69.0
60	65.0	71.0	72.2	70.6	78.4
75	67.2	75.8	80.2	79.6	81.4
90	77.0	80.8	84.8	81.2	83.4
100	77.0	83.4	87.4	86.2	85.6

We can see that as the reporting probability parameters increase, so does the power associated with each opportunity size. We expect this phenomenon to be the case because a higher parameter value indicates a greater probability of adverse events being reported. For instance, if the prior for p is a Beta(90,10), we can expect an average of $90/(90 + 10) = 90\%$ of adverse events to be reported. On the other hand, if the prior for p is a Beta(10,10), we expect only an average of $10/(10 + 10) = 50\%$ of adverse events to be reported. If all of the adverse events are reported, our results are similar to those of Model 1, in which we do not account for underreporting. However, since we do not always have complete reporting of adverse reactions, our results vary slightly.

We also note that as our opportunity size increases within each reporting probability prior, our power increases as well. For instance, the column with prior parameters 50 and 10 indicate a mean reporting of adverse events to be 50/(50+10) = 83%. The power greatly increases as the opportunity size increases from 50 to 100.

4.3.3 Varying Priors – Without Underreporting

A common prior for β would be independent normal distributions with zero mean and a common large variance. While a simple 'non-informative' choice is appealing, we would argue that some care is necessary. If the variance is too large, the induced prior on β could cause problems in our analysis. Thus, some care is necessary in the choice of the variance in order to obtain a reasonable 'non-informative' normal prior for the regression coefficients. We will take a look at different prior variances and how the results are affected.

We will look at three different prior distributions for the model not accounting for underreporting: highly diffuse, moderately difuse, and informative priors. For the highly diffuse priors, we use $\beta_0 \sim \text{Normal}(10, 50)$, $\beta_1 \sim \text{Normal}(3, 15)$, $\gamma_0 \sim \text{Normal}(0.1, 10)$, $\gamma_1 \sim \text{Normal}(\log(2), 10)$ and $\gamma_2 \sim \text{Normal}(0.1, 10)$.

Figure 4.3 displays the history plots for β_0, β_1, γ_0, γ_1 and γ_2 when the prior variances were highly diffuse. We can see that we have a lot of variability within the data. To account for this, we ran the same data with a thinning rate of three. After a thinning rate of three, there was still some variability, so we increased the thinning to ten. The results are in Figure 4.4. By thinning, we were able to account for the variability and autocorrelation.

For the moderately diffuse priors, we use $\beta_0 \sim \text{Normal}(10, 25)$, $\beta_1 \sim \text{Normal}(3, 10)$, $\gamma_0 \sim \text{Normal}(0.1, 5)$, $\gamma_1 \sim \text{Normal}(\log(2), 5)$ and $\gamma_2 \sim \text{Normal}(0.1, 5)$. We reached similar results with the moderately diffuse priors as we did with the highly diffuse priors. We adjusted the thinning rate to 10, and achieved more stability in the plots.

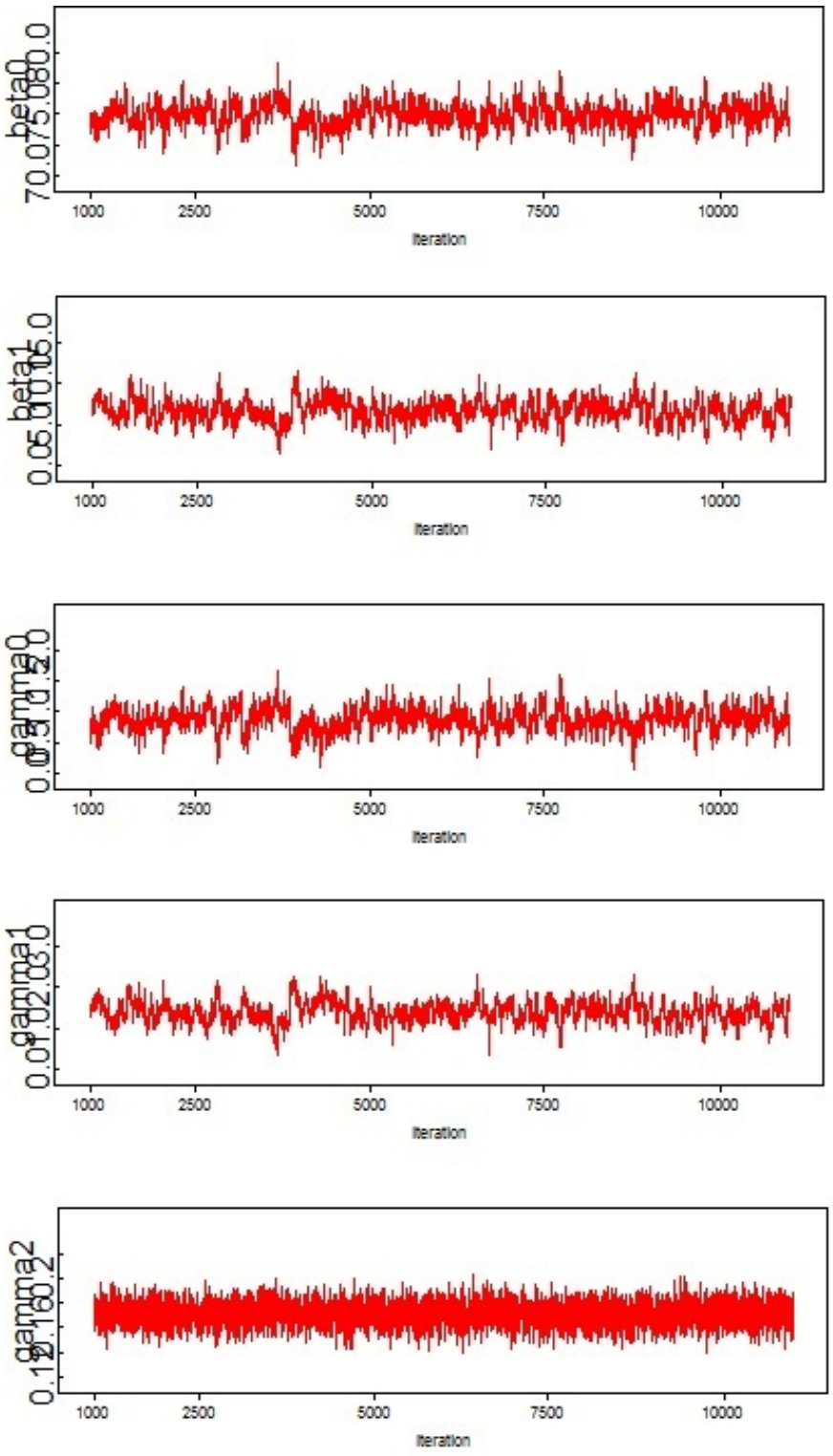

Figure 4.3: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with highly diffuse prior.

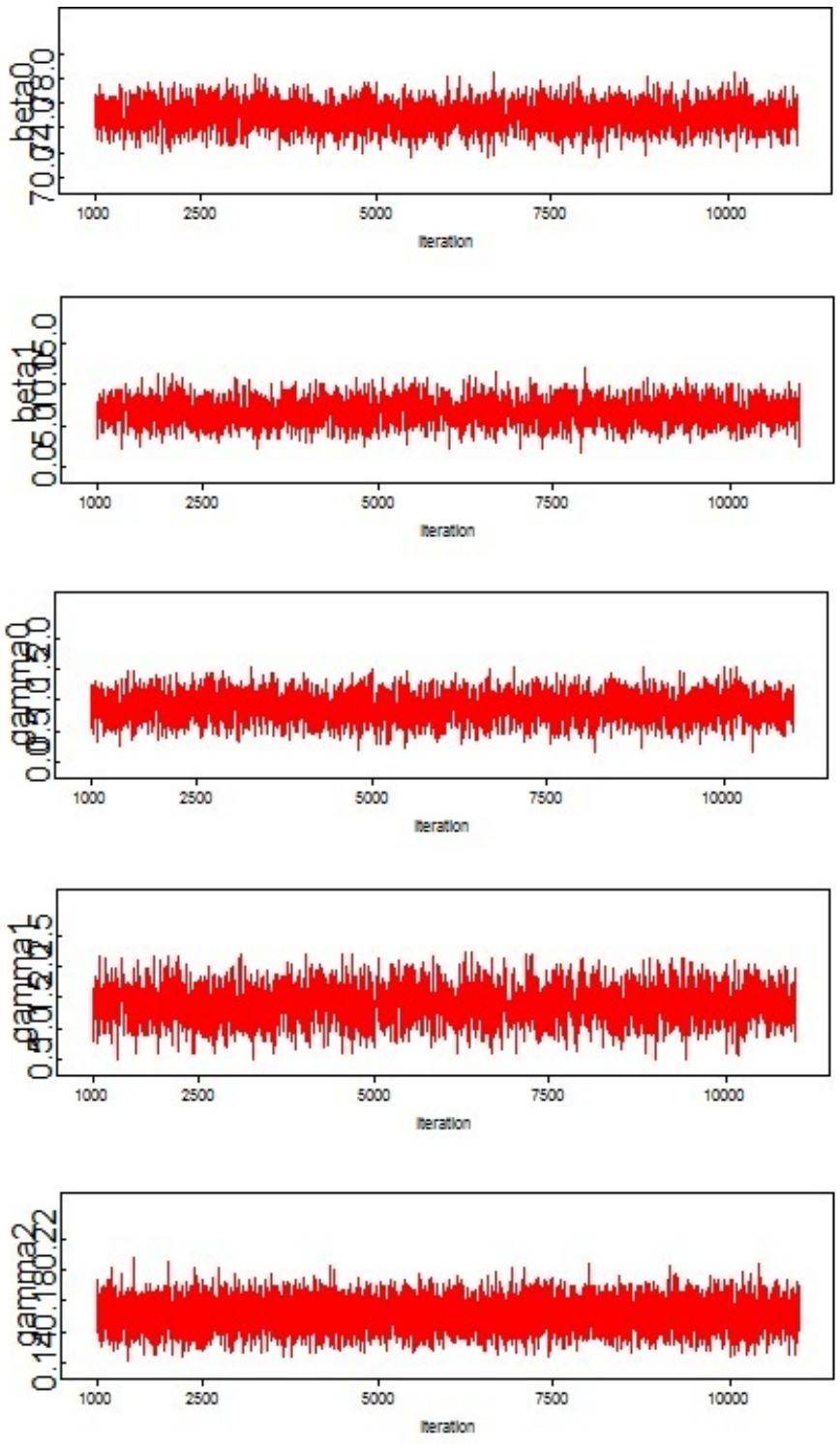

Figure 4.4: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with highly diffuse prior and thinning of 10.

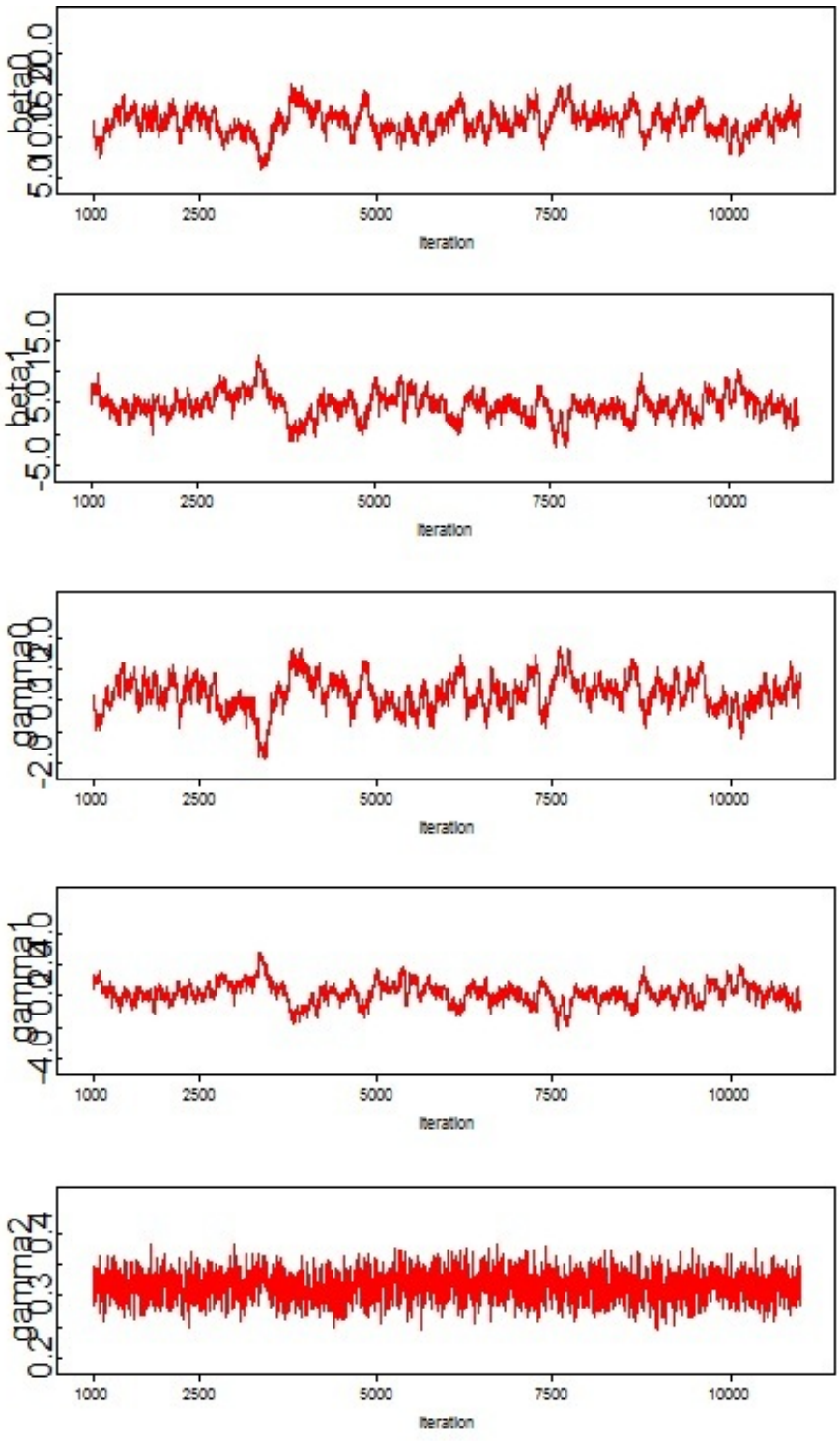

Figure 4.5: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with informative prior.

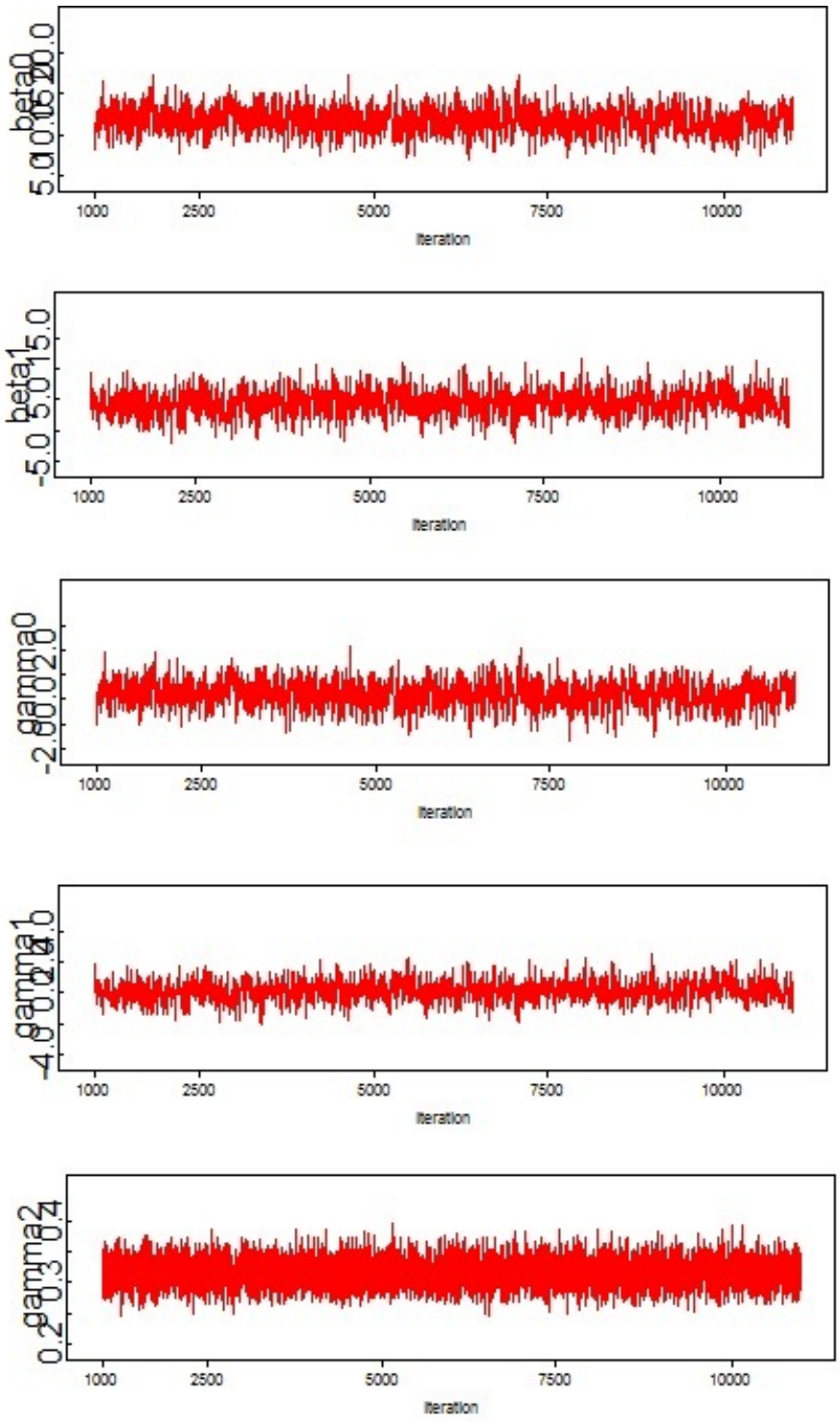

Figure 4.6: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with informative prior and thinning of 10.

For the informative priors, we use $\beta_0 \sim \text{Normal}(10, 2)$, $\beta_1 \sim \text{Normal}(3, 1)$, $\gamma_0 \sim \text{Normal}(0.1, 0.5)$, $\gamma_1 \sim \text{Normal}(\log(2), 1)$ and $\gamma_2 \sim \text{Normal}(0.1, 0.5)$. Figure 4.5 displays the history plots for β_0, β_1, γ_0, γ_1 and γ_2 when the prior variances are informative. Again, we have autocorrelation, so we thin. In this case, we thinned at a rate of ten. Figure 4.6 displays the results, which show more stability.

4.3.4 Varying Priors – With Underreporting

Now we will look at three different priors distributions for the model accounting for underreporting. For the highly diffuse priors, we use $\beta_0 \sim \text{Normal}(10, 50)$, $\beta_1 \sim \text{Normal}(3, 15)$, $\gamma_0 \sim \text{Normal}(0.1, 10)$, $\gamma_1 \sim \text{Normal}(\log(2), 10)$ and $\gamma_2 \sim \text{Normal}(0.1, 10)$. Because we have variability at the beginning of the chain, we adjust our burn-in (increased from 1000 to 5000) and thinning rate (thinned at ten). We still had some problems, so we changed our priors to slightly more informative. We now have the following highly diffuse priors: $\beta_0 \sim \text{Normal}(10, 5)$, $\beta_1 \sim \text{Normal}(3, 5)$, $\gamma_0 \sim \text{Normal}(0.1, 5)$, $\gamma_1 \sim \text{Normal}(\log(2), 5)$ and $\gamma_2 \sim \text{Normal}(0.1, 5)$.

For the moderately diffuse priors, we use $\beta_0 \sim \text{Normal}(10, 25)$, $\beta_1 \sim \text{Normal}(3, 10)$, $\gamma_0 \sim \text{Normal}(0.1, 5)$, $\gamma_1 \sim \text{Normal}(\log(2), 5)$ and $\gamma_2 \sim \text{Normal}(0.1, 5)$. The history plots are displayed in Figure 4.7. We had similar problems as with the highly diffuse priors. Thus, we changed our priors to be slightly more informative: $\beta_0 \sim \text{Normal}(10, 4)$, $\beta_1 \sim \text{Normal}(3, 3)$, $\gamma_0 \sim \text{Normal}(0.1, 2)$, $\gamma_1 \sim \text{Normal}(\log(2), 3)$ and $\gamma_2 \sim \text{Normal}(0.1, 2)$. Again, we changed our burn-in to 5000 and thinned at 10. Results are in Figure 4.8. Our plots look much more stable with these changes.

For the informative priors, we use $\beta_0 \sim \text{Normal}(10, 2)$, $\beta_1 \sim \text{Normal}(3, 1)$, $\gamma_0 \sim \text{Normal}(0.1, 0.5)$, $\gamma_1 \sim \text{Normal}(\log(2), 1)$ and $\gamma_2 \sim \text{Normal}(0.1, 0.5)$. Figure 4.9 displays the history plots for β_0, β_1, γ_0, γ_1 and γ_2 when the prior variances are informative. Again, we changed our burn-in and thinning to account for the variability in the plots. Figure 4.10 displays the changes, which again show stability.

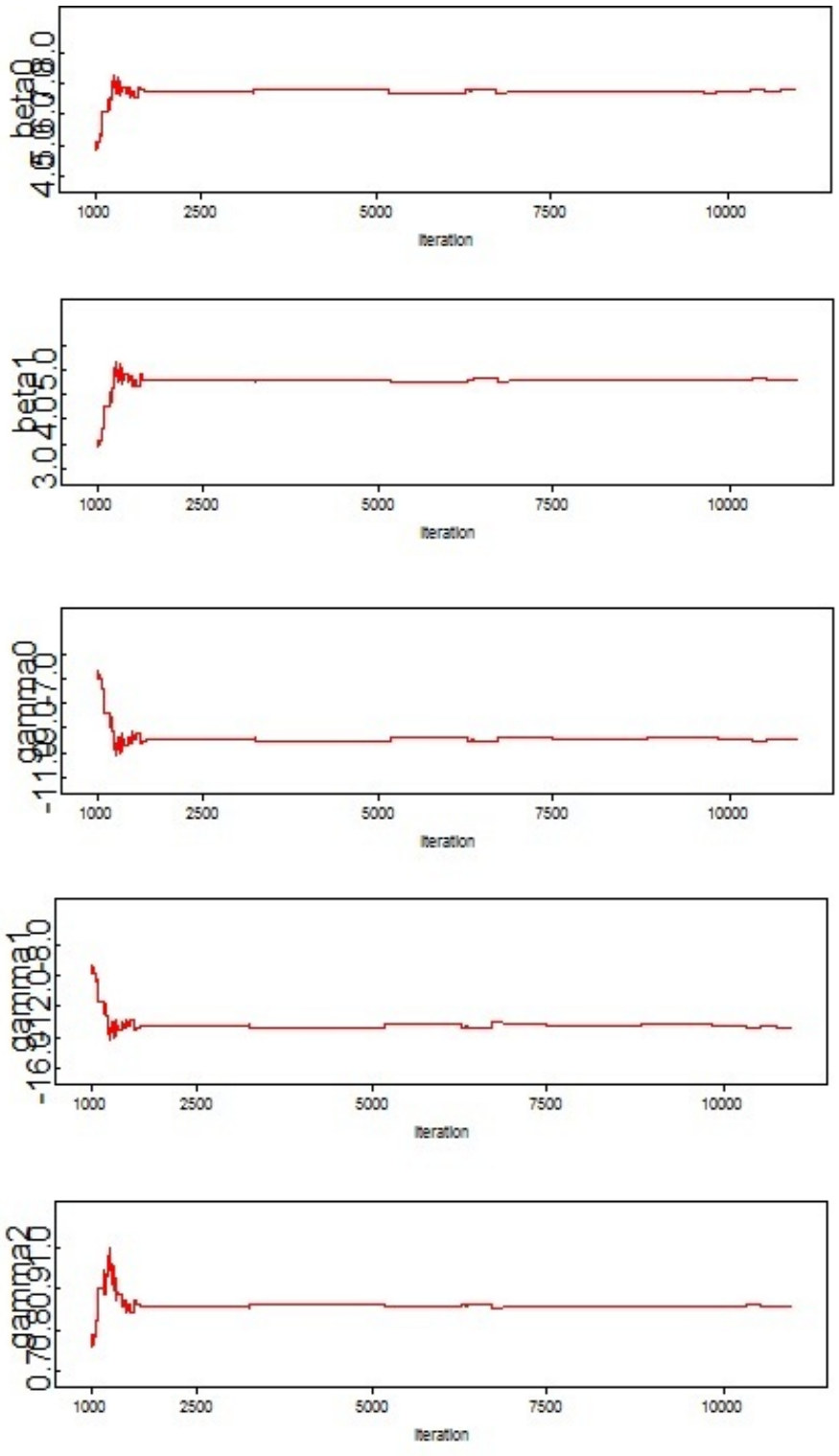

Figure 4.7: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with moderately diffuse prior.

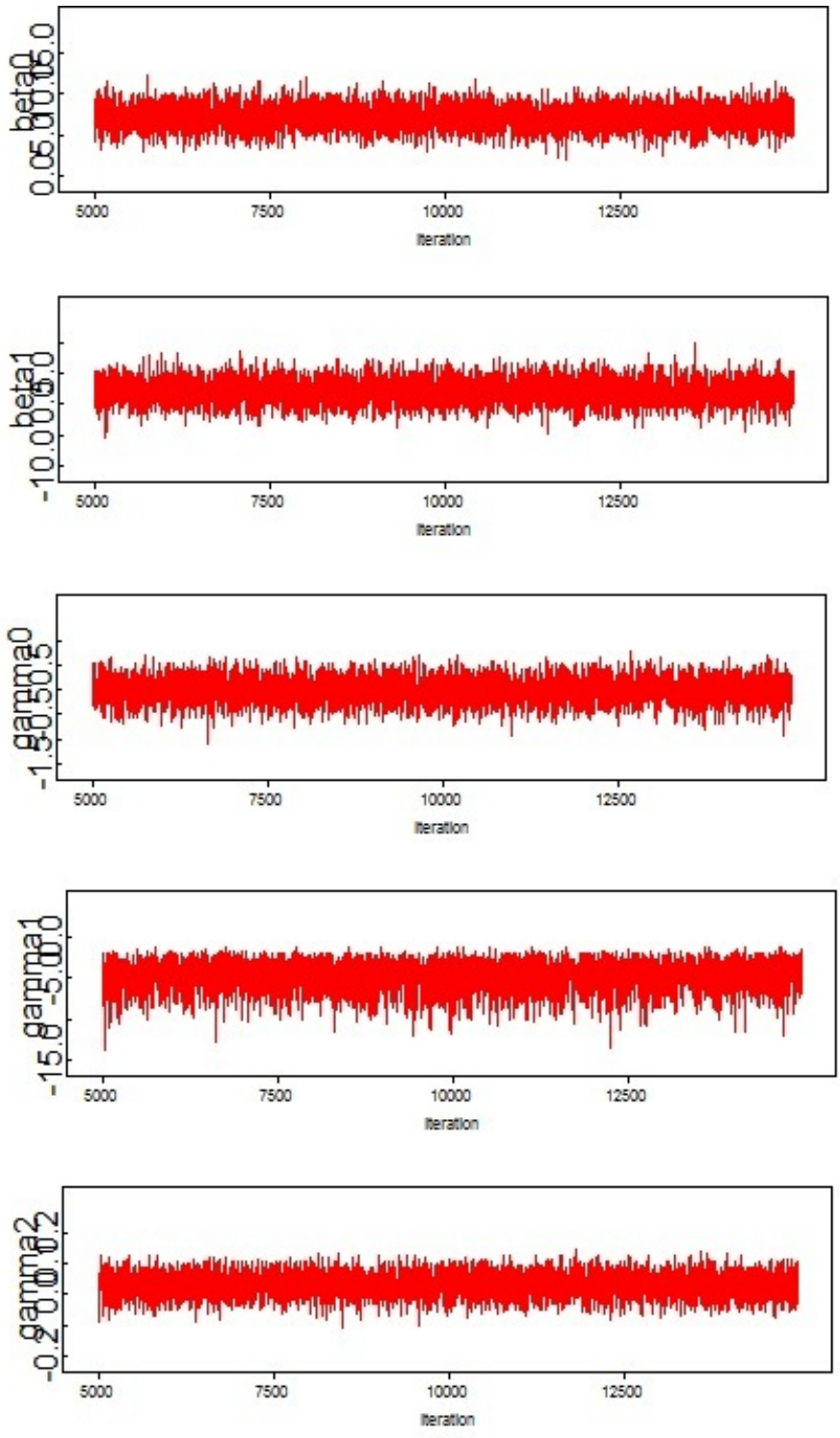

Figure 4.8: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with moderately diffuse prior, thinning of 10 and burn-in of 5000.

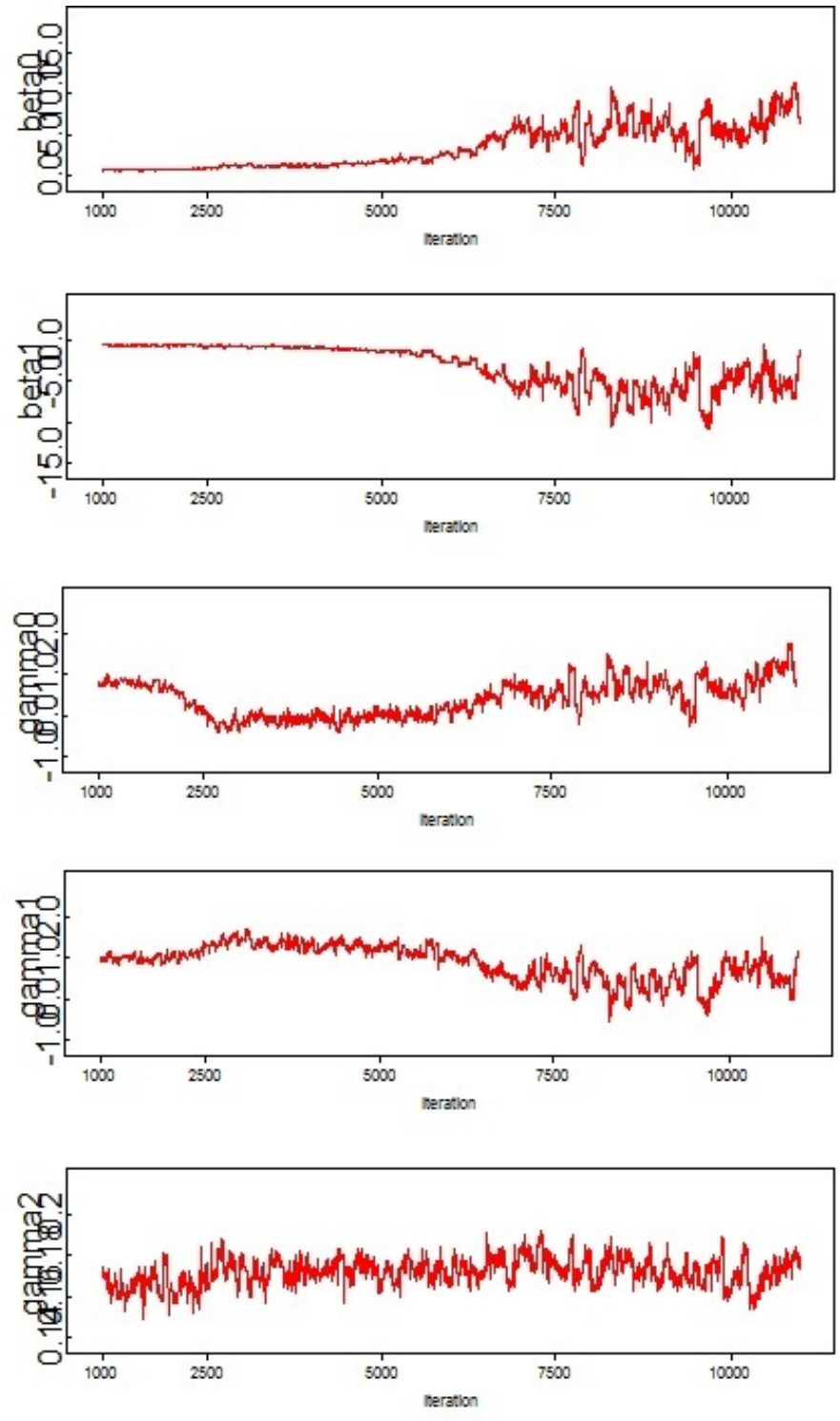

Figure 4.9: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with informative prior.

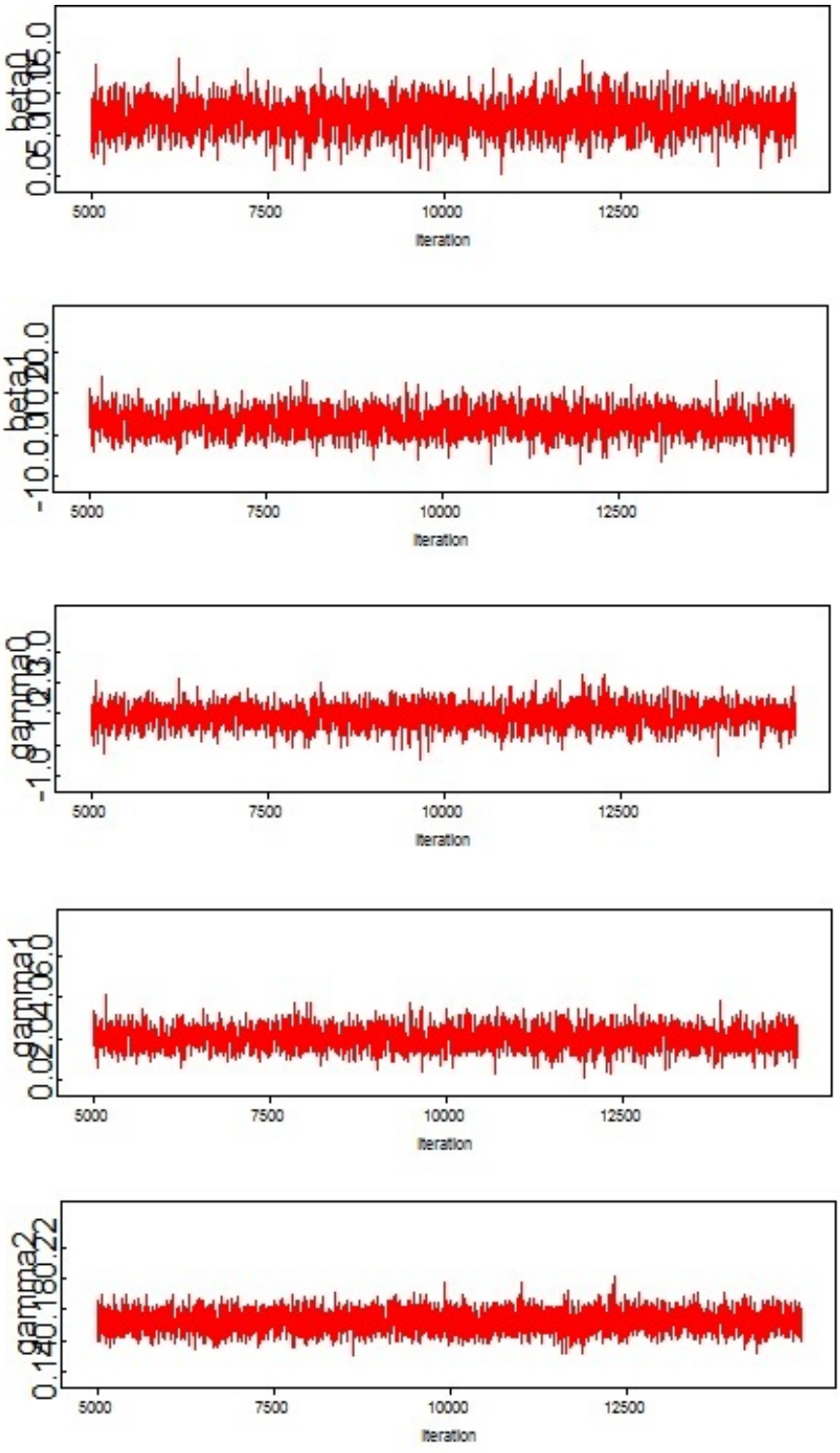

Figure 4.10: Convergence plots for β_0, β_1, γ_0, γ_1 and γ_2 for one data set with informative prior, thinning of 10 and burn-in of 5000.

4.3.5 Type I Error Analysis

We are interested in determining the required sample size to show that a parameter is "significantly" different from 0; a threshold other than 0 would be straightforward to incorporate into the procedure. If a positive relationship between the difference of the normally distributed responses is expected, a Bayesian power criterion selects n so that, for probabilities α and η,

$$E\left[I\left\{Pr\left(\beta_1 > 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta \tag{4.3}$$

where $I\{\}$, is the indicator function and $\mathbf{d^{(n)}}$ represents a generated data set of size n. Common choices for α are 0.1, 0.05, and 0.01 while η is typically 0.8 or 0.9. Here, the expectation is with respect to the design prior and the posterior probability, which is the argument of the indicator function. For each data set, $\mathbf{d^{(n)}}$, the null hypothesis of $H_0 : \beta_1 \leq 0$ is rejected in favor of $H_1 : \beta_1 > 0$ if $Pr(\beta_1 > 0|\mathbf{d^{(n)}}) > 1-\alpha$. With equation (4.3), we seek the sample size for which such hypotheses are rejected at least $100(1-\eta)\%$ of the time. If the relationship with the covariate is expected to be negative, equation (4.3) becomes

$$E\left[I\left\{Pr\left(\beta_1 < 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta.$$

Results are displayed in Figure 4.11(left).

If interest is in the treatment/safety relationship, a similar criterion is used. If a positive relationship between the difference of the treatments is expected, a Bayesian power criterion selects t so that, for probabilities α and η,

$$E\left[I\left\{Pr\left(\gamma_1 > 0|\mathbf{d^{(t)}}\right) > 1-\alpha\right\}\right] \geq \eta \tag{4.4}$$

where $\mathbf{d^{(t)}}$ represents a generated data set of size t. Figure 4.11(right) displays these results.

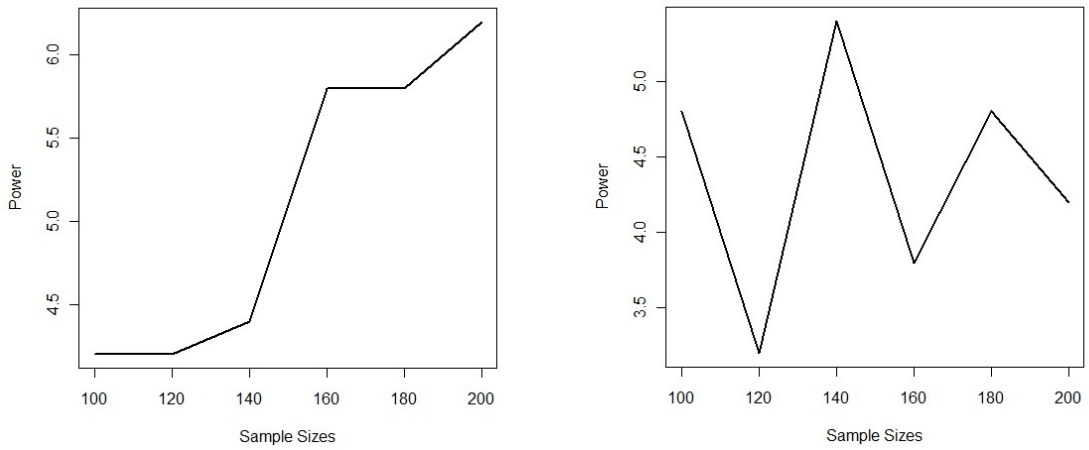

Figure 4.11: Type I error analysis results for β_1 when $\beta_1 = 0$ (left) and for γ_1 when $\gamma_1 = 0$ for the model not accounting for underreporting .

These procedures can be extended to the situation of testing multiple hypotheses simultaneously. For instance, testing both $H_0 : \beta_1 = 0$ and $H_0 : \gamma_1 = 0$ simultaneously may be of interest. Extension can be made to other multiple hypothesis structures as well. The goals of multiple testing need to be carefully considered. For instance, Sozu et al. (2011) provides a sample size formula for testing superiority with multiple end-points. In their case, the overall null hypothesis is rejected if each of the individual components is rejected and sample size is determined to achieve certain power to reject the overall null hypothesis. In our case, this would correspond to

$$E\left[I\left\{Pr\left(\beta_1 > 0|\mathbf{d^{(n)}}\right) > 1-\alpha \cap Pr\left(\gamma_1 > 0|\mathbf{d^{(n)}}\right) > 1-\alpha\right\}\right] \geq \eta \qquad (4.5)$$

This has a large impact on Type II error, but Type I error will be quite small using this framework. Figure **??** displays these results. An alternative would be to find the sample sizes for both criteria (4.3) and (4.4) and choose the larger. However, this approach tends to inflate Type I error, so the value of α may have to be adjusted slightly. Results of this multiple testing are in Figure 4.12.

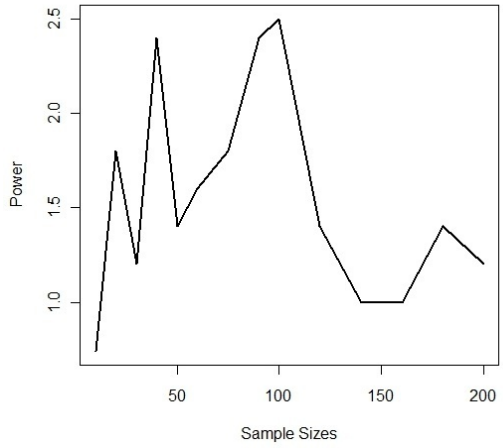

Figure 4.12: Type I error analysis results for the model not accounting for underreporting when $\delta_\mu = 0$ and $\delta_\lambda = 0$.

4.4 Discussion

Our models accurately estimate the difference in the means of the data as well as the difference in the Poisson rates. The model with underreporting was found to increase the necessary sample size.

In general, a tendency may exist for more active drugs to be both more efficacious and to cause more adverse reactions, leading to dependent prior distributions. An alternative way to relate the two variables is to incorporate a correlated random effect into the model. However, we feel this model is appropriate as often times the more effective the drug, the higher the dose, and thus leads to more adverse events. If this assumption is not reasonable, the random effects model would be straight forward to use.

The next step with these two models is to include covariates. This should be fairly straightforward due to the regression aspect of the model.

Due to the computational requirements, Monte Carlo or some other method of posterior approximation is required to estimate the sample size. We have considered the effect of prior information about p on the required sample size. Gains are more pronounced as the reporting probability nears 1. Finally, note that in some

applications, the sample sizes are in terms of linear feet and time, not a number of observations. For these applications the sample would not need to be constrained to an integer.

CHAPTER FIVE

A Meta Analysis

In this chapter, I will discuss a meta analysis project I worked on as an intern at Baylor Health Care Systems in Dallas, Texas. I worked with a heart surgeon, Dr. James Edgerton, who works at The Heart Hospital Baylor Plano.

5.1 Introduction

Atrial fibrillation (AF) is defined as supraventricular tachyarrhythmia characterized by the uncoordinated activation and deterioration of mechanical function of the atria (Fuster et al. (2006)). In the general population, the estimated prevalence of AF is 0.4%-1% (Go et al. (2001)), as AF is the most common cardiac arrhythmia encountered in clinical practice. There are a few options for the management of AF, which include pharmacologically achieved rate control, mechanistic prevention of thromboembolism, and the correction of the rhythm abnormality through surgical or catheter-based approaches (Fuster et al. (2006)).

Dr. Cox and colleagues initiated atrial fibrillation ablation with the introduction of the Maze procedure (2000). Although the Maze procedure has a high success rate (Edgerton and Edgerton (2009)) the operation required sternotomy access and arrest of the heart on cardiopulmonary bypass. The associated morbidities and complex nature of the procedure resulted in a relatively low adoption rate (Mack (2009); Edgerton and Edgerton (2009)). The Cox Maze III is rarely performed as a stand-alone procedure for AF, though it is still widely performed for AF concomitant to another cardiac surgical procedure (Ederton and Edgerton (2008)). For the treatment of stand-alone AF, surgery has largely been replaced by ever improving catheter based techniques. Consequently, catheter ablation is known to cause en-

docardial trauma, therapeutic thrombus, and high re-ablation rates due to limited efficacy (Cui et al. (2010)).

More recently, the development of enabling technology has allowed surgical ablation to be performed on the beating heart with techniques that require minimal access. This eliminates much of the morbidity associated with the original cut and sew Cox Maze III procedure. These minimally invasive surgical techniques hold the promise of higher potential curative benefits over catheter ablation (CA) for stand-alone AF, but have been only minimally discussed in the academic literature. Therefore, the present chapter describes the available publications from 2009 to 2011 of surgical intervention for stand-alone AF and provides an early depiction of summative results via preliminary meta-analysis. Results are then compared to published meta-analyses of CA success rates, and treatment recommendations are drawn.

5.2 Methods

5.2.1 Search Strategy

A comprehensive literature search was performed using the United States National Library of Medicine and the National Institutes of Health PubMed engine. The search criteria included all English manuscripts of observational studies of human subjects published from January 1, 2009, to September 10, 2011, with specified terms (See Appendix D).

All studies were examined to meet the inclusion criteria of longitudinal evaluation of freedom from atrial fibrillation (AF), atrial tachycardia (AT), or atrial flutter (Aflutter) following stand-alone ablative surgery with at least three months of follow-up data. Studies were excluded if information was published as an abstract, review, or case report. In addition, if the focus of the publication was a description of a surgical technique, or if the primary outcome of interest was different than free-

dom from AF or return to normal sinus rhythm (NSR), studies were excluded. A total of 1,364 published manuscripts were initially identified by MeSH key words in the identified time period. Figure 5.1 displays the exclusion criteria used to limit our analysis to the 13 remaining articles.

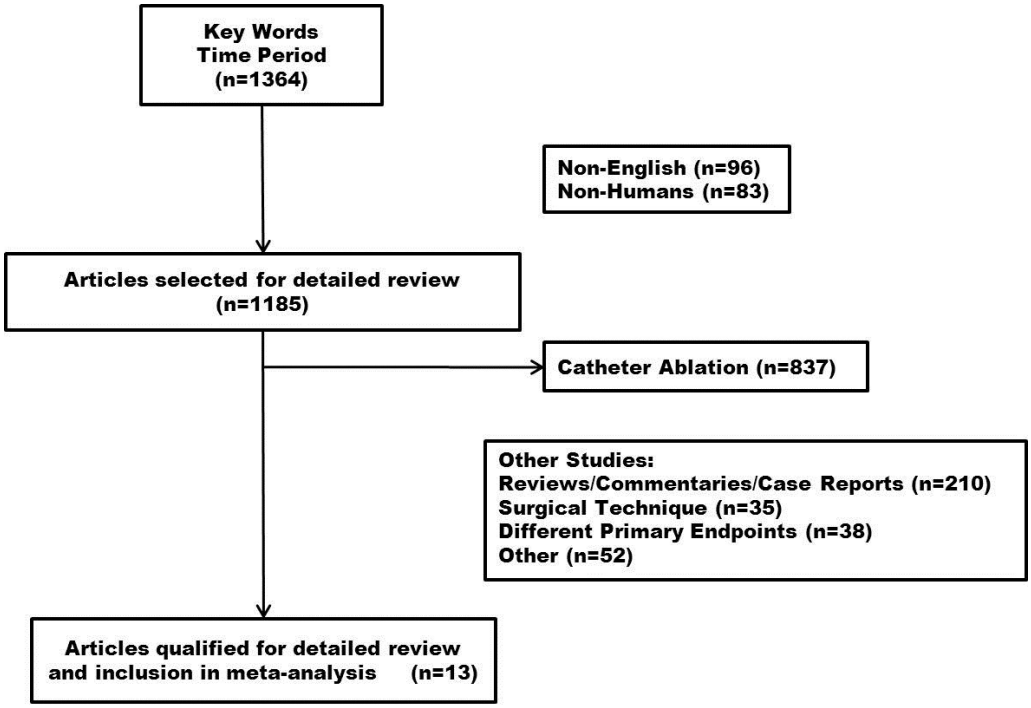

Figure 5.1: Schematic breakdown of studies included within systematic review and meta-analysis of stand-alone surgical ablation for atrial fibrillation (AF), 2009-2011.

5.2.2 Data Analysis

In each of the remaining articles, the following data elements were noted: total size of the study population, subset who received stand-alone ablative surgery (if applicable), mean age of study population, length of follow-up time in months, surgical technique used, primary outcome of interest, outcome assessment method, percent of patients who had previously undergone CA, mean left atrial size, and ejection fraction, if available. Type of AF was assigned according to the terminology recommended in the "Heart Rhythm Society (HRS)/European Heart Rhythm Association (EHRA)/European Cardiac Arrhythmia Society (ECAS) Expert Con-

sensus Statement on Catheter and Surgical Ablation of Atrial Fibrillation," in which paroxysmal AF is referred to as recurrent AF that terminates spontaneously within 7 days; persistent AF is AF sustained longer than 7 days or lasting less than 7 days but requiring either pharmacologic or electrical cardioversion; and long-standing persistent (LSP) AF is continuous AF of more than 1 year in duration.

5.2.3 Meta Analysis

Meta-analysis results are commonly displayed graphically as 'forest plots'. At a glance, forest plots show the effect size of all the studies, and the results of the meta-analysis.

Heterogeneity measures the variability between studies. In other words, it gives an indication how comparable studies in the meta-analysis are. Visually, we can assess heterogeneity by checking for overlap of the confidence intervals (CI). Studies are regarded as homogeneous if the CIs of all studies overlap.

Additionally, a test for heterogeneity examines the null hypothesis that all studies are evaluating the same effect. The usual test statistic, Cochrane's Q, is computed by summing the squared deviations of each study's estimate from the overall meta-analytic estimate, weighting each study's contribution in the same manner as in the meta-analysis:

$$Q = \sum w(E - E_C)^2,$$

where $E_C = \sum wE / \sum w$, w is the weight, and E is the effect size of the individual study. P values are obtained by comparing the statistic with a χ^2 distribution with $k - 1$ degrees of freedom (where k is the number of studies).

An alternative approach that quantifies the effect of heterogeneity, providing a measure of the degree of inconsistency in the studies' results, is the I^2 statistic. This quantity describes the percentage of total variation across studies that is due to heterogeneity rather than chance. I^2 can be readily calculated from basic results

obtained from a typical meta-analysis as

$$I^2 = 100\% \times \frac{(Q - df)}{Q},$$

where Q is Cochrane's heterogeneity statistic and df is the degrees of freedom. Negative values of I^2 are set equal to zero so that I^2 lies between 0% and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity.

The primary outcome was post-operative freedom from AF, AT, and Aflutter. All analyses were performed by stratifying included studies based on the described primary outcome: freedom from AF or return to NSR, according to the definitions provided within each manuscript. Estimates and pooled outcomes with 95% confidence intervals were calculated using fixed effects models. Statistical heterogeneity between studies was tested with the Cochrane test. The I^2 statistic was also examined, and $I^2 > 50\%$ was considered to signify heterogeneity between studies.

Publication bias was assessed via funnel plots. In a funnel plot we expect larger studies to be near the average and smaller studies to be spread on both sides of the average. As the studies become less precise (i.e. higher standard error), we expect the results of the studies to be more variable, scattered to both sides of the more precise larger studies. Variation from this assumption can indicate publication bias and this is seen in a funnel plot that shows an asymmetrical shape. Once the studies are plotted, if the plot is not symmetrical and does not resemble an inverted funnel, publication bias may be the cause. However, other factors lead to an assymetrical plot as well. If publication bias is not present, we expect the funnel plot to be roughly symmetrical. Statistical analyses were performed with R (version 2.13.1), and Appendix D contains the code for this chapter.

5.3 Results

Following all of the listed inclusion criteria, there were 13 published observational studies that included longitudinal follow-up of patients who underwent stand-alone ablative surgery for AF. In total, the 13 included articles allowed for a total study population of 699 patients. Six studies described freedom from AF as the primary endpoint, and 7 designated return to NSR as the primary endpoint. All included studies measured recurrence of AF or return to NSR through 24 hour-Holter monitor, electrocardiogram (ECG), AF monitor device, or a combination of the three. Table 5.1 displays these results.

Surgical technique varied among the studies: 8 used PVI, 2 studies used pulmonary vein isolation (PVI) with the Dallas Lesion Set (Edgerton et al. (2009a)), and 3 performed a version of the Cox Maze procedure. Report of previous CA based therapy varied greatly between studies, see Table 5.2.

5.3.1 Meta Analysis

The 13 described studies were then combined within a meta-analysis, the results of which can be seen in Figures 5.2 and 5.3. Combined results of postoperative freedom from AF (Figure 5.2A) indicate an overall 84% (80.0 to 88.0) success rate for the 6 included studies. Combined results of postoperative return to NSR indicate an 83% (79.0 to 87.0) success rate (Figure 5.2B). Cochrane evaluation of these groups indicated no heterogeneity, with I^2 values $< 50\%$ ($p > 0.05$).

Because of small sample sizes, the studies had to be combined to include both freedom from AF and return to NSR when stratified by type of AF (paroxysmal, persistent, or long-standing persistent). Those patients with paroxysmal AF who underwent stand-alone ablative surgery experienced an 85% (80.0 to 89.0) success rate when the study results were combined (Figure 5.3A). Patients with persistent AF had a 79% (0.70 to 86.0) success proportion (Figure 5.3B), but results are unstable

according to the Cochrane evaluation ($I^2 = 64.8\%$; $p = 0.0092$). The patients with long-standing persistent AF had a 64% (54.0 to 73.0) success rate in freedom from AF/return to NSR according to the 7 studies included in the analysis (Figure 5.3C).

The funnel plot in Figure 5.4 indicates a potential for publication bias within the study group, as the included studies do not represent a symmetrical pattern about the mean.

5.3.2 Results of Catheter Ablation Review Articles, Registries, Meta Analyses

To compare the surgical meta-analysis to catheter meta-analyses, we chose three meta-analyses of catheter ablation already published in peer reviewed literature. Each of these studies sought to collate published literature on the effectiveness and safety of at least one catheter-based therapy approach and utilized sound research methodology.

Calkins et al. (2009) published two separate systematic reviews and one meta-analysis in one manuscript, only a portion of which is applicable to the current discussion. In all, this review included 9 randomized clinical trials (RCTs) and 54 observational studies (42 prospective, 12 retrospective) with a total of 8,789 patients. The minimum follow-up time for inclusion was set at 7 days, and the study observed a mean age at time of ablative procedure of 55 years and duration of AF at 6.0 years. Patients with paroxysmal AF made up 35.8% of study participants, 33.3% had persistent AF, and 30.8% had long-standing persistent AF. Success was defined as the lack of recurrence of arrhythmia throughout the follow-up period, and the combined results indicated success following a single-procedure of patients either on or off anti-arrhythmic drug (AAD) therapy at 72%. For patients requiring multiple procedures, the success rate was 77% (73 to 81) following radiofrequency catheter ablation (RFA) either on or off AADs. The mean follow-up period for these studies was 14 months, with a range of 2 to 30 months.

In 2010, Kong et al. (2011) published a meta-analysis of 6 RCTs following a single catheter ablative procedure, with success defined as freedom from AF or AT, either with or without the ongoing use of AADs. All patients within this study received either pulmonary vein isolation (PVI) CA or PVI with complex fractionated atrial electrogram (CFAE) CA, and had a follow-up time of at least 3 months. A total of 538 patients were included within the analysis, 50.3% with paroxysmal AF and 49.7% with persistent AF. Combined results were stratified by length of follow-up, with 48% of patients who underwent PVI alone and 66% who underwent PVI+CFAE were free from AF/AT after one procedure with or without AADs, with a mean follow-up of 10.5 $\pm$ 1.8 months. Longer term follow-up results (14.2 $\pm$ 4.9 months) indicated 68% success following PVI alone and 82% success following PVI+CFAE respectively.

Li et al. (2011) performed and published a meta-analysis of seven controlled clinical trials (four with randomization, three without randomization) of patients following PVI+CFAE for AF. The authors sought to understand how CFAE following a single PVI procedure impacted the maintenance of sinus rhythm. In all, 662 patients were included within the analysis, with follow-up time ranging from 12 to 19 months. Patients with paroxysmal AF who underwent PVI experienced 75% maintenance of sinus rhythm, and 45% of patients with persistent or long-standing persistent AF achieved sinus rhythm maintenance.

Finally, we make reference to a systematic review of catheter ablation for long standing persistent atrial fibrillation, Brooks et al. (2010). When PVI alone was performed the success rate is 21%. When substrate ablation was added to PVI the mean success jumps to 47%.

Table 5.1: Descriptive Characteristics of Studies Included In Meta-Analysis of Lone Surgical Ablation for Atrial Fibrillation, 2009-2011.

Reference	Total N	Included N	Age (Mean ± SD)	Follow-Up (Months, Min*)	Outcome	Assessment Method
Albage et al. (2011)	43	6	54.2	12	NSR	ECG + Holter
Bagge et al. (2009)	43	33	57.1	12	Freedom from AF	Holter
Beukema et al. (2009)	33	33	59.4 ± 8.9	3	NSR	ECG + AF alarm
Beyer et al. (2009)	100	100	65 ± 11	3	NSR	Holter
Cui et al. (2010)	80	49	57.6 ± 10	12	NSR	ECG + Holter + UCG
Edgerton et al. (2009c)	114	114	59.5 ± 10	6	NSR	ECG + AF monitor
Edgerton et al. (2009b)	30	30	58 ± 9	6	Freedom from AF	ECG + AF monitor
Edgerton et al. (2010)	52	52	60.3	12	NSR	ECG + Holter
Kron et al. (2009)	50	37	63.4 ± 9.3	12	NSR	ECG + AF monitor
Krul et al. (2011)	31	22	57	12	Freedom from AF	ECG + Holter
Stulak et al. (2011)	289	93	56†	60	Freedom from AF	ECG + Holter
Weimar et al. (2011)	100	100	56 ± 10	24	Freedom from AF	ECG + Holter
Yilmaz et al. (2010)	30	30	55.6 ± 8.6	3	Freedom from AF	ECG + Holter

‡ Median reported; Abbreviations: AF = atrial fibrillation; ECG = electrocardiogram; Min = Minimum; NSR = normal sinus rhythm; UCG = Ultrasonic Cardiography.

Table 5.2. Available Study Characteristics Included In Meta-Analysis of Lone Surgical Ablation for Atrial Fibrillation, 2009-2011.

Reference	Surgical Technique	Patients with Previous CA (%)	Left Atrial Size Mean ± SD (mm)	Ejection Fraction (%)
Albage et al. (2011)	Cox III/IV	†	†	100% had EF > 50
Bagge et al. (2009)	PVI	†	45.0‡	86% had EF > 50
Beukema et al. (2009)	PVI	†	41.2	54 ± 4.3
Beyer et al. (2009)	PVI	100	43.0	55 ± 8.5
Cui et al. (2010)	PVI	4.9	49.7	†
Edgerton et al. (2009c)	PVI	21.1	47.2	50% had EF > 55
Edgerton et al. (2009b)	PVI + Dallas	†	52.0	†
Edgerton et al. (2010)	PVI	20.8	48.0	54.2
Kron et al. (2009)	PVI	†	†	54.9 ± 8.2
Krul et al. (2011)	PVI + Dallas	45	47.0	†
Stulak et al. (2011)	Cox III/IV	6.2	†	†
Weimar et al. (2011)	Cox III/IV	40	47.0	†
Yilmaz et al. (2010)	PVI	60	42.1	90% had EF > 40

† Information not provided; ‡ Median reported; Abbreviations: CA = Catheter Ablation; PVI = pulmonary vein isolation.

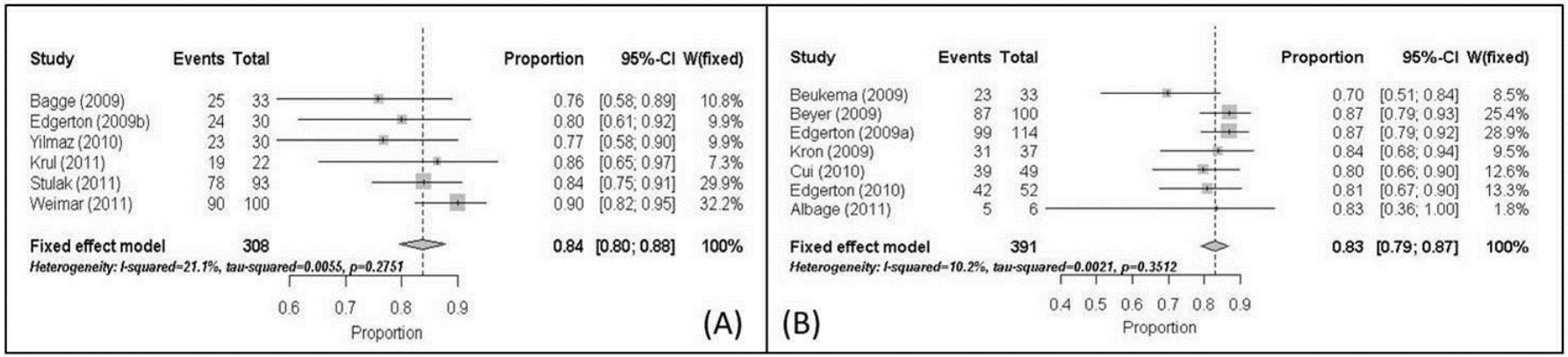

Figure 5.2: Forest Plots of combined studies of Freedom from AF (A) and Return to Normal Sinus Rhythm (B) following stand-alone surgical ablation for atrial fibrillation (AF), 2009-2011.

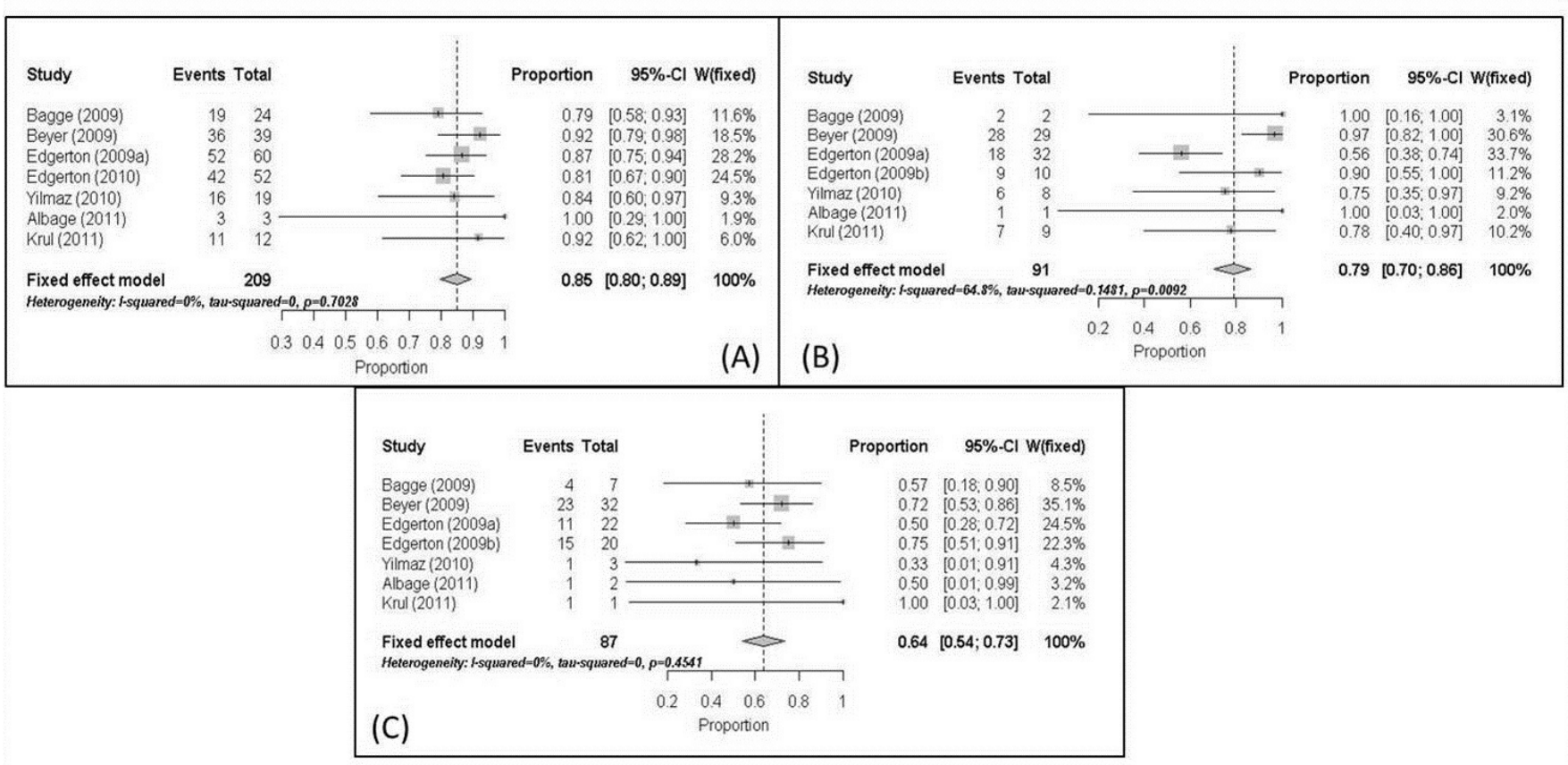

Figure 5.3: Forest Plots of combined studies of Paroxysmal AF (A), Persistent AF (B), and Long-Standing Persistent AF (C) following stand-alone surgical ablation, 2009-2011.

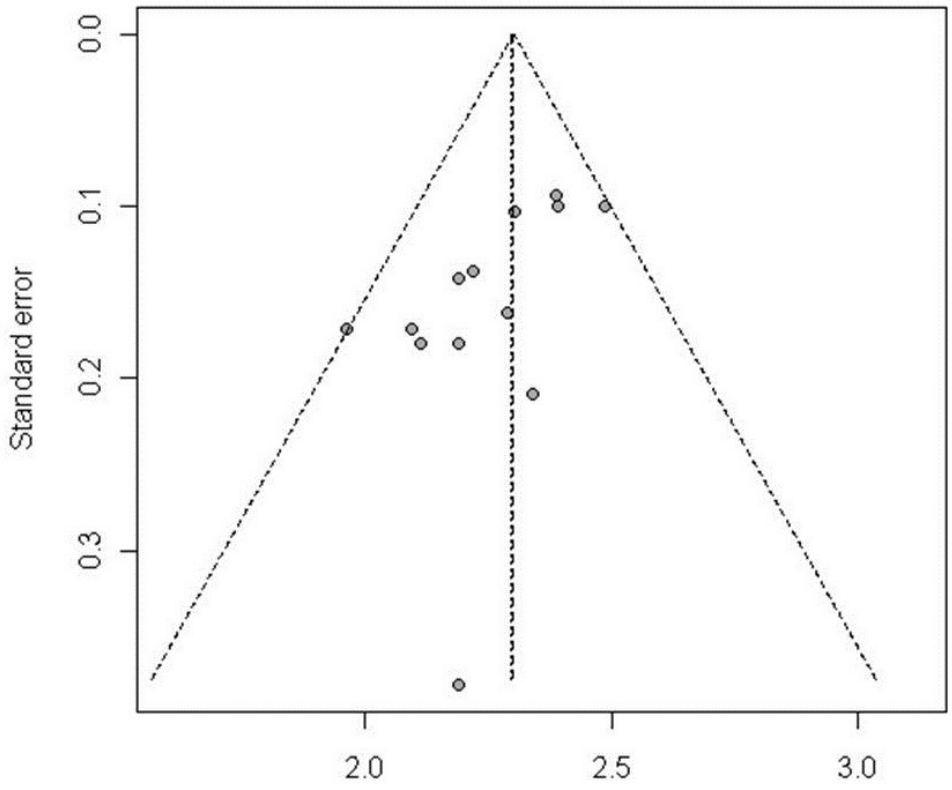

Figure 5.4: Funnel plot of all studies included (n=13) within systematic review and meta-analysis of stand-alone surgical ablation for atrial fibrillation (AF), 2009-2011.

5.4 Conclusions

The best available data from meta-analyses of catheter and surgical literature yields the following results.

The results from stand-alone surgical ablation are as follows. For overall success: post-operative freedom from AF with or without AAD is 84%, while post-operative freedom of AF, AT, Aflutter with or without AADs is 83%. Looking at success by AF type, we have: paroxysmal at 85%, persistent at 79%, and long standing persistent at 64% success.

The results from the catheter ablation literature are as follows. In Kong et al. (2011), freedom from AF/AT after one procedure with or without AADs using PVI is 48%, while using PVI+CFAE is 66%. Freedom from AF/AT after all procedures with or without AADs using PVI is 68%, while PVI+CFAE is 82%. In Calkins et al. (2009), freedom from AF without AADs in a single procedure is at 57%, while in multiple procedures, freedom from AF without AADs is 71%. Li et al. (2011) found paroxysmal to be at 75% and persistent/LSP to be at 45.25%. Finally, Brooks et al. (2010) noted freedom from AF/AT for LSP at 47%.

It seems clear to us that the initial ablative approach to patients with paroxysmal AF should be catheter ablation. The results of catheter ablation in this population approach those of surgery, with less morbidity. The paroxysmal patients can be treated with PVI alone, without the need to perform the more difficult linear ablations needed for substrate modification in more advanced types of AF. Additionally, improvements in catheter design and technology should further facilitate antral ablation.

For patients with persistent atrial fibrillation, there is insufficient evidence to argue for either catheter or surgical ablation. The ablative approach to these patients should be individualized with the more difficult patients being referred for minimal access surgical approach.

Based on the data above, we believe that the appropriate initial ablative procedure for patients with long standing persistent AF should be a totally thorascopic surgical Maze procedure. The surgical meta-analysis above shows a success rate of 64% contrasted with a 47% success for catheter ablation.

5.5 Discussion

Catheter ablation has made great strides in the treatment of atrial fibrillation. The results in paroxysmal AF are admirable and approach those of surgical ablation. Limitations of catheter ablation technology make it difficult to reliably produce the linear transmural lesions that are usually required for success in more advanced types of AF. One of the strengths of minimal-access surgical ablation is its ability to produce linear lesions. This contributes to the higher success rate in long-standing persistent AF and justifies surgery as the ablative procedure of choice for these difficult cases.

Readers should note that the potential for bias in estimation is inherent in the meta-analysis design. This analytic approach allows for the collation of a large number of studies and thereby increased sample size and power, but results should be interpreted with caution. Publication bias and the potential for heterogeneity of included studies can lead to bias estimation of success proportions. We utilized a funnel plot (Figure 5.4) as a tool to understand the potential for heterogeneity of study populations, which indicate a need for caution in interpretation of results. This heterogeneity may be caused by differing definitions of AF and success. Additionally, not all studies reported complete information on proportion by type of AF and defining characteristics of assessed outcomes. This lack of information does not allow for assured collation of outcomes within the meta-analysis.

Practitioners in the future will be encouraged or required to work in multidisciplinary teams. Those who dispense with old competitive predispositions and adopt

this new paradigm will be most successful. The precedent is set for this collaboration. The Society of Thoracic Surgeons, American College of Cardiology, the Food and Drug Administration, and the Center for Medicare and Medicaid Services have joined together to collaboratively introduce transcatheter aortic valve replacement as a mandatory multidisciplinary team approach with mandatory long term follow-up (Mack and Holmes (2011)). More work is needed in all these areas of ablation of atrial fibrillation. Future studies should report their results in compliance with the HRS, EHRA, ECAS expert consensus statement (Calkins et al. (2007)) so that results can be easily compared and conclusions more clearly made.

CHAPTER SIX

Conclusion

6.1 Future Work in the Area of Misclassification

The main advantage of a Bayesian approach for measurement error problems is that it allows the problem to be modeled in a conceptually straightforward way without approximations (Ren and Stone (2007)). All the available information is utilized and the uncertainty from different sources is properly reflected in the parameter estimates. Moreover, it works under more complicated model frameworks and accounts for measurement error in a relatively straightforward way.

Our methods in Chapter Two incorporated the information from two dichotomous test, one continuous test, and the misclassification of the covariates, using logistic regression, into a single model, while not considering any method to be a gold standard. We have improved on the previous work by combining all these tests into one model, while allowing the covariates to be misclassified. In particular, we do not require that these parameters be known and provide the means to incorporate information about them from previous studies and expert opinion. We use prior distributions to model our uncertainty about the values of the parameters in both the response model and the measurement error model. In this way, the Bayesian context provides an attractive method for adjusting inferences to account for measurement error and misclassification. Given that there is no gold standard, our credible intervals do not appear wide, indicating little uncertainty about the prevalence of a disease.

As for the Math data set, we need a much larger sample size. Our data set only contained 181 subjects, which accounts for the large standard deviations in the results. A larger sample size in our math data set should provide better results.

6.2 Future Work in the Area of Independent Efficacy and Safety Sample Size Determination

In Chapter Three, we determined the necessary sample size for independent efficacy and safety. We compared the model with underreporting to the model without underreporting. Our models accurately estimate the difference in the means of the data as well as the difference in the Poisson rates. The model with underreporting was found to increase the needed sample size.

In general, there may be a tendency for more active drugs to be both more efficacious and to cause more adverse events, leading to dependent prior distributions. However, there are also many cases for which an assumption of independence is reasonable. Published examples include calcineurin inhibitors for immunosuppression in liver transplantation Perry and Neuberger (2005), rosuvastatin to reduce low-density lipoprotein cholesterol Olsson et al. (2001), Davidson et al. (2002), Saito et al. (2003), and infliximab (a monoclonal antibody against Tumor Necrosis Factor) for Crohn's disease Targan et al. (1997).

6.3 Future Work in the Area of Dependent Efficacy and Safety Sample Size Determination

In Chapter Four, we determined the necessary sample size for efficacy and safety, while allowing the two variables to be dependent through regression modeling. Again, we compared the model with underreporting to the model without underreporting. Our models accurately estimate the difference in the means of the data and the difference in the Poisson rates, and the model with underreporting was found to increase the necessary sample size.

An alternative way to relate efficacy and safety is to incorporate a correlated random effect into the model. However, we feel our model is appropriate as often times the more effective the drug, the higher the dose, which leads to more adverse

events. If this assumption is not reasonable, the random effects model would be straight forward to use.

6.4 Future Work in the Area of Atrial Fibrillation

Chapter Five provides the cumulative results of a preliminary meta-analysis of the available stand alone atrial fibrillation surgical intervention publications from 2009 to 2011. The primary outcome of this review was postoperative freedom from atrial fibrillation, atrial tachycardia, and atrial flutter. After inclusion and exclusion criteria were applied, there were 13 eligible published observational studies that included longitudinal follow-up of patients undergoing stand-alone surgical ablation for AF.

For patients with persistent atrial fibrillation, there was insufficient evidence to argue for either catheter or surgical ablation. The ablative approach to these patients should be individualized with the more difficult patients being referred for minimal access surgical approach. We believe that the appropriate initial ablative procedure for patients with long standing persistent AF should be a totally thorascopic surgical Maze. The surgical meta-analysis shows a success rate of 64% contrasted with a 47% success for catheter ablation.

Practitioners in the future will be encouraged or required to work in multidisciplinary teams. Those who dispense with old competitive predispositions and adopt this new paradigm will be most successful. The precedent is set for this collaboration. The Society of Thoracic Surgeons, American College of Cardiology, the Food and Drug Administration, and the Center for Medicare and Medicaid Services have joined together to collaboratively introduce transcatheter aortic valve replacement as a mandatory multidisciplinary team approach with mandatory long term follow-up (Mack and Holmes (2011)). More work is needed in all these areas of ablation of atrial fibrillation. Future studies should report their results in compliance with

the HRS, EHRA, ECAS expert consensus statement (Calkins et al. (2007)) so that results can be easily compared and conclusions more clearly made.

APPENDICES

APPENDIX A

Misclassification Codes

A.1 R Code

```
library(R2WinBUGS)
library(xtable)
m=500
bugs.out=vector()
data.sens=matrix(0,m,4)
data.spec=matrix(0,m,4)
data.beta0=matrix(0,m,4)
data.beta1=matrix(0,m,4)
data.sens.X=matrix(0,m,4)
data.spec.X=matrix(0,m,4)
data.deviance=matrix(0,m,4)
coverage.sens=vector()
coverage.spec=vector()
coverage.beta0=vector()
coverage.beta1=vector()
coverage.sens.X=vector()
coverage.spec.X=vector()


for(i in 1:m){
N=1000
   p.E = 0.5
   beta0 = -2
```

```
   beta1 = 0.5
#original values
   sens.X = 0.8 #0.9
   spec.X = 0.9 #0.8
#alternative values (.9/.7)
   sens = 0.9 #0.7
   spec = 0.7 #0.9
#E is the exposure, the true status of covariates
#defines the unobserved true status data
E <- rbinom(N, 1,p.E)
#Bern prob for fallible test
p.X<- E*sens.X + (1-E)*(1-spec.X)
#fallible diagnostic test for exposure, E
#defines the observed data sent to WinBUGS
X <- rbinom(N, 1, p.X)
continuous.test.mean <- c(38.4, 118.6)
continuous.test.sd <- c(21.3, 48.6)
continuous.test.mean.mean=c(0, 0)
continuous.test.mean.tau=c(.001, .001)
continuous.test.sd.lower=c(0.1, 0.1)
continuous.test.sd.upper=c(60, 120)
b1.mu=0
b1.prec=.1
b0.mu=0
b0.prec=.1
sens.alpha.x = 80
sens.beta.x = 20
```

```
spec.alpha.x = 90

spec.beta.x = 10

#original (centered) values for priors: (for sens/spec = .9/.7)

#for sens/spec = .7/.9

sens.alpha = 90 #70

sens.beta=10 #30

spec.alpha = 70 #90

spec.beta=30 #10

#alternative (offset) values for priors: (for sens/spec = .9/.7)

#for sens/spec = .7/.9

sens.alpha = 84 #76

sens.beta=16 #24

spec.alpha = 76 #84

spec.beta=24 #16

#logit(prev) <- beta0 + beta1*E

prev <- exp(beta0 + beta1*E) / (1 + exp(beta0 + beta1*E))

true.status <- rbinom(N, 1,prev)

true.status.index <- 1 + true.status

#T is a dicohtomous test and p.T is the probability of

       #a positive test result

p.T <- sens * true.status + (1-spec)*(1 - true.status)

T <- rbinom(N,1,p.T)

continuous.test <-

  rnorm(N, continuous.test.mean[true.status.index],

  continuous.test.sd[true.status.index])


#the data list from the defined data
```

```
data.list<- list("E", "T", "continuous.test", "N",
  "continuous.test.mean.mean", "continuous.test.mean.tau",
  "continuous.test.sd.lower", "continuous.test.sd.upper",
  "continuous.test", "b1.mu", "b1.prec", "b0.mu", "b0.prec",
  "sens.alpha", "sens.beta", "spec.alpha", "spec.beta",
  "sens.alpha.x", "sens.beta.x", "spec.alpha.x",
  "spec.beta.x", "true.status")
#parameters of interest
parameters<-list("sens","spec", "beta0", "beta1",
  "sens.X", "spec.X")
#initial values
initials<-list(list(beta0=-2, beta1=0.5, sens=0.9, spec=0.8,
sens.X = 0.8, spec.X = 0.9, p.E=0.5,
continuous.test.mean=c(38.4, 118.6),
continuous.test.sd=c(21.3, 48.6),
true.status=
c(0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1,
1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0,
1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0,
1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1,
0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1,
1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1,
1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1,
1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1,
```

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,

0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0,

1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1,

1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,

1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1,

0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0,

1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1,

1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,

0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1,

0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1,

0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1,

1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1,

1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0,

0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0,

1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1,

1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1,

1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1,

1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,

1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0,

0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1,

1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0,

0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1,

0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1,

0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1,

0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,

1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1,

0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,

```
0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1,
1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0,
0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1,
0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1,
0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1,
1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0,
1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1,
1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0)))
#calling WinBUGS, given the initial values and parameters above
posterior =bugs(data=data.list, inits=initials, parameters,
  "Bugs.txt", n.iter=5000, n.chains=1,n.burnin=1000,
  n.thin=3, debug=TRUE)

#[,1] mean, [,2] sd, [,3] 2.5%, [,7] 97.5%
data.sens[i,]=c(posterior$summary[1,1],posterior$summary[1,2],
  posterior$summary[1,3],posterior$summary[1,7])
data.spec[i,]=c(posterior$summary[2,1],posterior$summary[2,2],
  posterior$summary[2,3],posterior$summary[2,7])
data.beta0[i,]=c(posterior$summary[3,1],posterior$summary[3,2],
  posterior$summary[3,3],posterior$summary[3,7])
data.beta1[i,]=c(posterior$summary[4,1],posterior$summary[4,2],
  posterior$summary[4,3],posterior$summary[4,7])
data.sens.X[i,]=c(posterior$summary[5,1],posterior$summary[5,2],
  posterior$summary[5,3],posterior$summary[5,7])
data.spec.X[i,]=c(posterior$summary[6,1],posterior$summary[6,2],
```

```
  posterior$summary[6,3],posterior$summary[6,7])
data.deviance[i,]=c(posterior$summary[7,1],posterior$summary[7,2],
  posterior$summary[7,3],posterior$summary[7,7])

############ Coverage ############
#sens
if(sens>posterior$summary[1,3] && sens<posterior$summary[1,7])
{ cov.sens=1 }
   else { cov.sens=0 }
coverage.sens=c(coverage.sens,cov.sens)
#spec
if(spec>posterior$summary[2,3] && spec<posterior$summary[2,7])
{ cov.spec=1 }
   else { cov.spec=0 }
coverage.spec=c(coverage.spec,cov.spec)
#beta0
if(beta0>posterior$summary[3,3] && beta0<posterior$summary[3,7])
{ cov.beta0=1 }
   else { cov.beta0=0 }
coverage.beta0=c(coverage.beta0,cov.beta0)
#beta1
if(beta1>posterior$summary[4,3] && beta1<posterior$summary[4,7])
{ cov.beta1=1 }
   else { cov.beta1=0 }
coverage.beta1=c(coverage.beta1,cov.beta1)
#sens.X
if(sens.X>posterior$summary[5,3] && sens.X<posterior$summary[5,7])
```

```
{ cov.sens.X=1 }
   else { cov.sens.X=0 }
coverage.sens.X=c(coverage.sens.X,cov.sens.X)
#spec.X
if(spec.X>posterior$summary[6,3] && spec.X<posterior$summary[6,7])
{ cov.spec.X=1 }
   else { cov.spec.X=0 }
coverage.spec.X=c(coverage.spec.X,cov.spec.X)
}
#print our bugs output by parameter
#  (average of all m simulations)
avg.sens = c(mean(data.sens[,1]), mean(data.sens[,2]),
  mean(data.sens[,3]), mean(data.sens[,4]))
avg.spec = c(mean(data.spec[,1]), mean(data.spec[,2]),
  mean(data.spec[,3]), mean(data.spec[,4]))
avg.beta0 = c(mean(data.beta0[,1]), mean(data.beta0[,2]),
  mean(data.beta0[,3]), mean(data.beta0[,4]))
avg.beta1 = c(mean(data.beta1[,1]), mean(data.beta1[,2]),
  mean(data.beta1[,3]), mean(data.beta1[,4]))
avg.sens.X = c(mean(data.sens.X[,1]), mean(data.sens.X[,2]),
  mean(data.sens.X[,3]), mean(data.sens.X[,4]))
avg.spec.X = c(mean(data.spec.X[,1]), mean(data.spec.X[,2]),
  mean(data.spec.X[,3]), mean(data.spec.X[,4]))
avg.deviance=c(mean(data.deviance[,1]),mean(data.deviance[,2]),
  mean(data.deviance[,3]), mean(data.deviance[,4]))

avg.cov.sens = sum(coverage.sens)/m
```

```
avg.cov.spec = sum(coverage.spec)/m
avg.cov.beta0 = sum(coverage.beta0)/m
avg.cov.beta1 = sum(coverage.beta1)/m
avg.cov.sens.X = sum(coverage.sens.X)/m
avg.cov.spec.X = sum(coverage.spec.X)/m
#print summary stats
#print coverage
```

A.2 WinBUGS Code, Simulated Data

```
model
{
# N: number of observations
for (i in 1:N)
{
#true.status = Y, 0 or 1
true.status[i] ~ dbern(prev[i])
true.status.index[i] <- 1 + true.status[i]
# -- Covariate --
#E is the exposure, the true status of covariates
logit(prev[i]) <- beta0 + beta1*E[i]
E[i] ~ dbern(p.E)

#X is the fallible diagnostic test for exposure
#X is the observed covariates
X[i] ~ dbern(p.X[i])
p.X[i] <- E[i]*sens.X+(1-E[i])*(1-spec.X)
#sens.X and spec.X are the sens and spec for X,
#the true status of the covariates
```

```
# -- Dichotomous test(s) --
#diagnostic tests are T (T_1, T_2, etc.), 0 or 1
T[i] ~ dbern(pos.diag.pr[i, 1, true.status.index[i]])

pos.diag.pr[i, 1, 2] <- sens
pos.diag.pr[i, 1, 1] <- 1 - spec
# -- Continuous test --
#continuous.test = W
continuous.test[i] ~
  dnorm(continuous.test.mean[true.status.index[i]],
  continuous.test.tau[true.status.index[i]])
}
# ---- Priors ----------------------------------
#p.X is the Bern prob for true status of covariates
p.X ~ dbeta(50, 50)

sens ~ dbeta(sens.alpha, sens.beta)
spec ~ dbeta(spec.alpha, spec.beta)
sens.X ~ dbeta(sens.alpha.x, sens.beta.x)
spec.X ~ dbeta(spec.alpha.x, spec.beta.x)
beta0 ~ dnorm(b0.mu, b0.prec)
beta1 ~ dnorm(b1.mu, b1.prec)

for (j in 1:2)
{
continuous.test.mean[j] ~ dnorm(continuous.test.mean.mean[j],
  continuous.test.mean.tau[j])
```

```
continuous.test.sd[j] ~ dunif(continuous.test.sd.lower[j],
  continuous.test.sd.upper[j])
continuous.test.tau[j] <- 1/pow(continuous.test.sd[j],2)
}
}
```

A.3 WinBUGS Code, Beaujean Data, With Misclassification

```
#with X and T both misclassified
model
{
# N: number of observations
for (i in 1:N)
{
#true.status = Y, 0 or 1
true.status[i] ~ dbern(prev[i])
true.status.index[i] <- 1 + true.status[i]
# -- Covariate --
#E is the exposure, the true status of covariates
logit(prev[i]) <- beta0 + beta1*E[i] + beta2*Z1[i]
  + beta3*Z2[i] + beta4*Z3[i]
E[i] ~ dbern(p.E[i])
logit(p.E[i]) <- gamma0 + gamma1*Z1[i]
  + gamma2*Z2[i] + gamma3*Z3[i]
#Z1 is location, Z2 is HSGPA, Z3 is sex

#X is the fallible diagnostic test for exposure;
# the observed covariates
#X is the verbal points
```

```
#Se1, Se2, Sp1, Sp2 are the sens and spec for the 2 X's,
#for the true status of the covariates
# 1,0,0,0 if both 1s; 0,1,0,0 if MJ3PC 1 and MJ3RF 0
# 0,0,1,0 if MJ3PC 0 and MJ3RF 1; 0,0,0,1 if both 0s
p.X[i,1] <- E[i]*se.X1*se.X2 + (1-E[i])*(1-sp.X1)*(1-sp.X2)
p.X[i,2] <- E[i]*se.X1*(1-se.X2) + (1-E[i])*(1-sp.X1)*sp.X2
p.X[i,3] <- E[i]*(1-se.X1)*se.X2 + (1-E[i])*sp.X1*(1-sp.X2)
p.X[i,4] <- E[i]*(1-se.X1)*(1-se.X2) + (1-E[i])*sp.X1*sp.X2
X[i,1:4] ~ dmulti(p.X[i,1:4],1)
# -- Dichotomous test(s) --
#D=2
#two math tests, MF, MCS
# 1,0,0,0 if both 1s; 0,1,0,0 if MF 1 and MCS 0
# 0,0,1,0 if MF 0 and MCS 1; 0,0,0,1 if both 0s

#diagnostic tests are T (T_1, T_2, etc.), 0 or 1
p.T[i,1] <- true.status[i]*se1*se2
  + (1-true.status[i])*(1-sp1)*(1-sp2)
p.T[i,2] <- true.status[i]*se1*(1-se2)
  + (1-true.status[i])*(1-sp1)*sp2
p.T[i,3] <- true.status[i]*(1-se1)*se2
  + (1-true.status[i])*sp1*(1-sp2)
p.T[i,4] <- true.status[i]*(1-se1)*(1-se2)
  + (1-true.status[i])*sp1*sp2

T[i,1:4] ~ dmulti(p.T[i,1:4],1)
# -- Continuous test --
```

```
#math, AP
#continuous.test = W
continuous.test[i] ~
  dnorm(continuous.test.mean[true.status.index[i]],
  continuous.test.tau[true.status.index[i]])
}
# ---- Priors ----------------------------------
#p.X is the Bern prob for true status of covariates
#p.X ~ dbeta(50, 50)

se1 ~ dbeta(sens.alpha, sens.beta)
sp1 ~ dbeta(spec.alpha, spec.beta)
se2 ~ dbeta(sens.alpha, sens.beta)
sp2 ~ dbeta(spec.alpha, spec.beta)
se.X1 ~ dbeta(sens.alpha.x, sens.beta.x)
sp.X1 ~ dbeta(spec.alpha.x, spec.beta.x)
se.X2 ~ dbeta(sens.alpha.x, sens.beta.x)
sp.X2 ~ dbeta(spec.alpha.x, spec.beta.x)
beta0 ~ dnorm(b0.mu, b0.prec)
beta1 ~ dnorm(b1.mu, b1.prec)
beta2 ~ dnorm(b2.mu, b2.prec)
beta3 ~ dnorm(b3.mu, b3.prec)
beta4 ~ dnorm(b4.mu, b4.prec)

for (j in 1:2)
{
continuous.test.mean[j] ~ dnorm(continuous.test.mean.mean[j],
```

```
  continuous.test.mean.tau[j])
continuous.test.sd[j] ~ dunif(continuous.test.sd.lower[j],
  continuous.test.sd.upper[j])
continuous.test.tau[j] <- 1/pow(continuous.test.sd[j],2)
}
}
```

A.4 WinBUGS Code, Beaujean Data, No Misclassification

```
# BUGS code, new
model
{
# N: number of observations
for (i in 1:N)
{
#true.status = Y, 0 or 1
true.status[i] ~ dbern(prev[i])
true.status.index[i] <- 1 + true.status[i]
# -- Covariate --
logit(prev[i]) <- beta0 + beta1*E[i] + beta2*Z1[i]
+ beta3*Z2[i] + beta4*Z3[i]
#E is the exposure, the true status of covariates
#Z1 is location, Z2 is HSGPA, Z3 is sex
#use MJ3PC (Passage Comprehension) for E
# -- Dichotomous test(s) --
T[i] ~ dbern(prev[i])
#use MCS for T
# -- Continuous test --
#math, AP
```

```
#continuous.test = W
continuous.test[i] ~
  dnorm(continuous.test.mean[true.status.index[i]],
  continuous.test.tau[true.status.index[i]])
}
# ---- Priors ----------------------------------
se ~ dbeta(sens.alpha, sens.beta)
sp ~ dbeta(spec.alpha, spec.beta)
se.X ~ dbeta(sens.alpha.x, sens.beta.x)
sp.X ~ dbeta(spec.alpha.x, spec.beta.x)
beta0 ~ dnorm(b0.mu, b0.prec)
beta1 ~ dnorm(b1.mu, b1.prec)
beta2 ~ dnorm(b2.mu, b2.prec)
beta3 ~ dnorm(b3.mu, b3.prec)
beta4 ~ dnorm(b4.mu, b4.prec)

for (j in 1:2)
{
continuous.test.mean[j] ~ dnorm(continuous.test.mean.mean[j],
  continuous.test.mean.tau[j])
continuous.test.sd[j] ~ dunif(continuous.test.sd.lower[j],
  continuous.test.sd.upper[j])
continuous.test.tau[j] <- 1/pow(continuous.test.sd[j],2)
}
}
```

APPENDIX B

Independent Efficacy and Safety Codes

B.1 R Code for the Model Not Accounting for Underreporting

```
library(R2WinBUGS)
library(MASS)
n<-100 #data: Normal - sample size
t<-100 #data: Poisson - sample size
m<-500


data.delta=matrix(0,m,4)
data.delta.r=matrix(0,m,4)
p.value1 <- vector()
p.value2 <- vector()
power.lambda=matrix(0,m,4)
power.mu=matrix(0,m,4)
for(k in 1:m){
alpha1 <- 4
   #shape parameter for lambda1 gamma for population 1
beta1 <- 1
   #scale parameter for lambda1 gamma for population 1
delta.r <- 2
lambda1 <- rgamma(1, alpha1, beta1)
   #mean: alpha1/beta1
lambda2 <- delta.r + lambda1
   #want this to be 2
```

```
mu.theta <- 80
   #mean for theta
tau.theta <- 5
   #variance for theta
delta <- 5
theta <- rnorm(1, mu.theta, tau.theta) #mean: 80
mean.x <- theta + delta
mu.r1 <- t*lambda1
mu.r2 <- t*lambda2
#safety
r1 <- rpois(1, mu.r1)
r2 <- rpois(1, mu.r2)
a=15
b=20
sig <- runif(1, a, b)
tau <- 1/(sig*sig)
#efficacy
y <- rnorm(n, theta, sig)
x <- rnorm(n, mean.x, sig)

#the data list from the defined data
data.list<-list("n","t","r1","r2","y","x")
#parameters of interest
parameters<-list("delta", "delta.r","p.val.lambda","p.val.mu")
#initial values
initials<-list(list(lambda1=1, theta=80 , sig=17))
#calling WinBUGS, given the initial values and parameters above
```

```
pow =bugs(data=data.list, inits=initials, parameters,
  "BUGS_Ch3_NOunderrep_07MAY12.txt", n.iter=10000,
  n.chains=1, n.burnin=1000)

#[,1] mean, [,2] sd, [,3] 2.5%, [,7] 97.5%
data.delta[k,]=c(pow$summary[1,1],pow$summary[1,2],
  pow$summary[1,3],pow$summary[1,7])
data.delta.r[k,]=c(pow$summary[2,1],pow$summary[2,2],
  pow$summary[2,3],pow$summary[2,7])
power.lambda[k,]=c(pow$summary[3,1],pow$summary[3,2],
  pow$summary[3,4],pow$summary[3,6])
power.mu[k,]=c(pow$summary[4,1],pow$summary[4,2],
  pow$summary[4,4],pow$summary[4,6])

# assign a value of 1 if the posterior probability
# is greater than 0.95 and a value of 0 otherwise
if(power.mu[k,1]>0.95){p.value1[k]=1} else p.value1[k]=0
if(power.lambda[k,1]>0.95){p.value2[k]=1} else p.value2[k]=0
}
avg.delta=c(mean(data.delta[,1]), mean(data.delta[,2]),
  mean(data.delta[,3]), mean(data.delta[,4]))
avg.delta
avg.delta.r=c(mean(data.delta.r[,1]), mean(data.delta.r[,2]),
  mean(data.delta.r[,3]), mean(data.delta.r[,4]))
avg.delta.r

#the average power
```

```
power.delta.mu <- mean(p.value1)

power.delta.lam <- mean(p.value2)

power.delta.mu

power.delta.lam
```

B.2 WinBUGS Code for the Model Not Accounting for Underreporting

```
model

{

for (i in 1:n)

{

#efficacy

y[i] ~ dnorm(theta, tau)

x[i] ~ dnorm(mean.x, tau)

}

theta ~ dnorm(0, 0.0001) #non-informative

mean.x ~ dnorm(0, 0.0001)

delta <- mean.x - theta

#mean.x <- theta + delta

#delta ~ dnorm(0, 0.000001)

tau <- 1/(sig*sig)

#sig ~ dunif(.01,200)

sig ~ dunif(.1,50)

#safety

r1 ~ dpois(mu.r1)

r2 ~ dpois(mu.r2)

mu.r1 <- t*lambda1

mu.r2 <- t*lambda2

lambda1 ~ dgamma(.01, .01) #non-informative
```

```
lambda2 ~ dgamma(.01, .01)
delta.r <- lambda2 - lambda1
#lambda2 <- delta.r + lambda1
#delta.r ~ dnorm(0, 0.000001)
p.val.lambda <- step(delta.r)
p.val.mu <- step(delta)
}
```

B.3 R Code for the Model Accounting for Underreporting

```
library(R2WinBUGS)
library(MASS)
n<-20 #data: Normal - sample size
t<-20 #data: Poisson - sample size
m<-500
data.delta=matrix(0,m,4)
data.delta.r=matrix(0,m,4)
p.value1 <- vector()
p.value2 <- vector()
power.lambda=matrix(0,m,4)
power.mu=matrix(0,m,4)

for(k in 1:m){
alpha1 <- 4 #shape parameter for lambda1 gamma
beta1 <- 1 #scale parameter for lambda1 gamma
delta.r <- 2
lambda1 <- rgamma(1, alpha1, beta1) #mean: alpha1/beta1
lambda2 <- delta.r + lambda1 #want this to be 2
mu.theta <- 80 #mean for theta
```

```
tau.theta <- 5 #variance for theta

delta <- 5

theta <- rnorm(1, mu.theta, tau.theta) #mean: 80

mean.x <- theta + delta

c <- 90 #first parameter for binomial prior

d <- 10 #first parameter for binomial prior

p <- rbeta(1, c, d)

A1 <- t*lambda1*p

A2 <- t*lambda2*p

#underreporting, safety

w1 <- rpois(1, A1)

w2 <- rpois(1, A2)

a=15

b=20

sig <- runif(1, a, b)

tau <- 1/(sig*sig)

#efficacy

y <- rnorm(n, theta, sig)

x <- rnorm(n, mean.x, sig)


#the data list from the defined data

data.list<-list("n","c","d","t","y","x","w1","w2")

#parameters of interest

parameters<-list("delta", "delta.r","p.val.lambda","p.val.mu")

#initial values

initials<-list(list(lambda1=1, theta=80 , sig=17 , p=.9))

#calling WinBUGS, given the initial values and parameters above
```

```
pow=bugs(data=data.list, inits=initials, parameters,
  "BUGS_Ch3_Underreporting_07MAY12.txt", n.iter=10000,
  n.chains=1, n.burnin=1000)

#[,1] mean, [,2] sd, [,3] 2.5%, [,7] 97.5%
data.delta[k,]=c(pow$summary[1,1],pow$summary[1,2],
  pow$summary[1,3],pow$summary[1,7])
data.delta.r[k,]=c(pow$summary[2,1],pow$summary[2,2],
  pow$summary[2,3],pow$summary[2,7])
power.lambda[k,]=c(pow$summary[3,1],pow$summary[3,2],
  pow$summary[3,4],pow$summary[3,6])
power.mu[k,]=c(pow$summary[4,1],pow$summary[4,2],
  pow$summary[4,4],pow$summary[4,6])

# assign a value of 1 if the posterior probability is greater
# than 0.95 and a value of 0 otherwise
if(power.mu[k,1]>0.95){p.value1[k]=1} else p.value1[k]=0
if(power.lambda[k,1]>0.95){p.value2[k]=1} else p.value2[k]=0
}

avg.delta = c(mean(data.delta[,1]), mean(data.delta[,2]),
  mean(data.delta[,3]), mean(data.delta[,4]))
avg.delta
avg.delta.r = c(mean(data.delta.r[,1]), mean(data.delta.r[,2]),
  mean(data.delta.r[,3]), mean(data.delta.r[,4]))
avg.delta.r
```

```
#the average power
power.delta.mu <- mean(p.value1)
power.delta.lam <- mean(p.value2)
power.delta.mu
power.delta.lam
```

B.4 WinBUGS Code for the Model Accounting for Underreporting

```
model
{
for (i in 1:n)
{
#efficacy
y[i] ~ dnorm(theta, tau)
x[i] ~ dnorm(mean.x, tau)
}
theta ~ dnorm(0, 0.0001) #non-informative
mean.x ~ dnorm(0, 0.0001)
delta <- mean.x - theta
#mean.x <- theta + delta
#delta ~ dnorm(0, 0.000001)
tau <- 1/(sig*sig)
#sig ~ dunif(.01,200)
sig ~ dunif(.1,50)
#underreporting for safety
w1 ~ dpois(A1)
w2 ~ dpois(A2)
A1 <- t*lambda1*p
A2 <- t*lambda2*p
```

```
p ~ dbeta(c, d)

lambda1 ~ dgamma(.01, .01) #non-informative

lambda2 ~ dgamma(.01, .01)

delta.r <- lambda2 - lambda1

#lambda2 <- delta.r + lambda1

#delta.r ~ dnorm(0, 0.000001)

p.val.lambda <- step(delta.r)

p.val.mu <- step(delta)

}
```

APPENDIX C

Code for Dependent Efficacy and Safety

C.1 R Code for the Model Not Accounting for Underreporting

```
library(R2WinBUGS)

library(MASS)

n<-165

m<-500

data.beta1=matrix(0,m,4)

data.gamma1=matrix(0,m,4)

power.beta=matrix(0,m,4)

power.gamma=matrix(0,m,4)

p.value1 <- vector()

p.value2 <- vector()


for(k in 1:m){

beta0 <- rnorm(1, 10, 3)

beta1 <- 3

#treatment has an increase of 'beta1' units

gamma0 <- rnorm(1, 0.1, 0.05)

gamma1 <- log(2) #difference in treatments is exp(gamma1)

gamma2 <- rnorm(1, 0.1, 0.05) #needs to be small

a=5

b=10

sig <- runif(1, a, b)

tau <- 1/(sig*sig)
```

```
Z <- rbinom(n, 1, 0.5)
theta <- beta0 + beta1*Z
#efficacy
y <- rnorm(n, theta, sig)
#safety
mu.r <- exp(gamma0 + gamma1*Z + gamma2*(y-theta))
r <- rpois(n, mu.r)

#the data list from the defined data
data.list<-list("n","r","y","Z")
#parameters of interest
parameters<-list("beta1","gamma1","p.val.beta","p.val.gamma")
#initial values
initials<-list(list(lambda1=1))
#calling WinBUGS, given the initial values and parameters above
pow =bugs(data=data.list, inits=initials, parameters,
  "BUGS_Ch4_NOunderrep.txt", n.iter=10000,
  n.chains=1, n.burnin=1000, n.thin=1)

#[,1] mean, [,2] sd, [,3] 2.5%, [,7] 97.5%
data.beta1[k,]=c(pow$summary[1,1],pow$summary[1,2],
  pow$summary[1,4],pow$summary[1,6])
data.gamma1[k,]=c(pow$summary[2,1],pow$summary[2,2],
  pow$summary[2,4],pow$summary[2,6])
power.beta[k,]=c(pow$summary[3,1],pow$summary[3,2],
  pow$summary[3,4],pow$summary[3,6])
power.gamma[k,]=c(pow$summary[4,1],pow$summary[4,2],
```

```
  pow$summary[4,4],pow$summary[4,6])

# assign a value of 1 if the posterior probability is
# greater than 0.95 and a value of 0 otherwise
if(power.beta[k,1]>0.95){p.value1[k]=1} else p.value1[k]=0
if(power.gamma[k,1]>0.95){p.value2[k]=1} else p.value2[k]=0
}
avg.beta1 = c(mean(data.beta1[,1]), mean(data.beta1[,2]),
  mean(data.beta1[,3]), mean(data.beta1[,4]))
avg.beta1
avg.gamma1 = c(mean(data.gamma1[,1]), mean(data.gamma1[,2]),
  mean(data.gamma1[,3]), mean(data.gamma1[,4]))
avg.gamma1

#the average power
power.beta1 <- mean(p.value1)
power.gamma1 <- mean(p.value2)
power.beta1
power.gamma1
```

C.2 WinBUGS Code for the Model Not Accounting for Underreporting

```
model
{
for (i in 1:n)
{
y[i] ~ dnorm(theta[i], tau)
#this includes all the data (x and y)
theta[i] <- beta0 + beta1*Z[i]
```

```
#Z is an indicator. if Z=0, data=y. if Z=1, data=x
r[i] ~ dpois(mu.r[i]) #all r's together
log(mu.r[i]) <- gamma0 + gamma1*Z[i] + gamma2*(y[i]-theta[i])
}
tau <- 1/(sig*sig)
sig ~ dunif(0.1, 50)
beta0 ~ dnorm(0, 0.0001)
beta1 ~ dnorm(0, 0.0001)
gamma0 ~ dnorm(0, 0.1)
gamma1 ~ dnorm(0, 0.1)
gamma2 ~ dnorm(0, 0.1)
p.val.beta <- step(beta1)
p.val.gamma <- step(gamma1)
}
```

C.3 R Code for the Model Accounting for Underreporting

```
library(R2WinBUGS)
library(MASS)
n<-180
m<-500
data.beta1=matrix(0,m,4)
data.gamma1=matrix(0,m,4)
power.beta=matrix(0,m,4)
power.gamma=matrix(0,m,4)
p.value1 <- vector()
p.value2 <- vector()

for(k in 1:m){
```

```
beta0 <- rnorm(1, 10, 3)
beta1 <- 3
#treatment has an increase of 'beta1' units
gamma0 <- rnorm(1, 0.1, 0.05)
gamma1 <- log(2) #difference in treatments is exp(gamma1)
gamma2 <- rnorm(1, 0.1, 0.05) #needs to be small
a=5
b=15
sig <- runif(1, a, b)
tau <- 1/(sig*sig)
Z <- rbinom(n, 1, 0.5)
theta <- beta0 + beta1*Z
#efficacy
y <- rnorm(n, theta, sig)
c <- 50 #first parameter for binomial prior
d <- 10 #first parameter for binomial prior
p <- rbeta(1, c, d)
#safety
#underreporting
mu.r <- exp(gamma0 + gamma1*Z + gamma2*(y-theta))
mu.w <- p*mu.r
w <- rpois(n, mu.w)

#the data list from the defined data
data.list<-list("n","w","y","Z")
#parameters of interest
parameters<-list("beta1","gamma1","p.val.beta","p.val.gamma")
```

```
#initial values
initials<-list(list(beta0=0, beta1=0, gamma0=0,
gamma1=0, gamma2=0, p=.9))
#calling WinBUGS, given the initial values and parameters above
pow =bugs(data=data.list, inits=initials, parameters,
  "BUGS_Ch4_Underreporting.txt", n.iter=10000,
  n.chains=1, n.burnin=1000, n.thin=10)

#[,1] mean, [,2] sd, [,3] 2.5%, [,7] 97.5%
data.beta1[k,]=c(pow$summary[1,1],pow$summary[1,2],
  pow$summary[1,3],pow$summary[1,7])
data.gamma1[k,]=c(pow$summary[2,1],pow$summary[2,2],
  pow$summary[2,3],pow$summary[2,7])
power.beta[k,]=c(pow$summary[3,1],pow$summary[3,2],
  pow$summary[3,4],pow$summary[3,6])
power.gamma[k,]=c(pow$summary[4,1],pow$summary[4,2],
  pow$summary[4,4],pow$summary[4,6])

# assign a value of 1 if the posterior probability is
# greater than 0.95 and a value of 0 otherwise
if(power.beta[k,1]>0.95){p.value1[k]=1} else p.value1[k]=0
if(power.gamma[k,1]>0.95){p.value2[k]=1} else p.value2[k]=0
}
avg.beta1 = c(mean(data.beta1[,1]), mean(data.beta1[,2]),
  mean(data.beta1[,3]), mean(data.beta1[,4]))
avg.beta1
avg.gamma1 = c(mean(data.gamma1[,1]), mean(data.gamma1[,2]),
```

```
  mean(data.gamma1[,3]), mean(data.gamma1[,4]))
avg.gamma1


#the average power
power.beta1 <- mean(p.value1)
power.gamma1 <- mean(p.value2)
power.beta1
power.gamma1
```

C.4 WinBUGS Code for the Model Accounting for Underreporting

```
model
{
for (i in 1:n)
{
y[i] ~ dnorm(mu.y[i], tau)
#this includes all the data (x and y)
mu.y[i] <- beta0 + beta1*Z[i]
#Z is an indicator. if Z=0, data=y. if Z=1, data=x
w[i] ~ dpois(mu.w[i]) #all r's together
log(mu.r[i]) <- gamma0 + gamma1*Z[i] + gamma2*(y[i]-mu.y[i])
mu.w[i] <- p*mu.r[i]
}
tau <- 1/(sig*sig)
sig ~ dunif(0.1, 50)
beta0 ~ dnorm(0, 0.01)
beta1 ~ dnorm(0, 0.01)
gamma0 ~ dnorm(0, 0.1)
gamma1 ~ dnorm(0, 0.1)
```

```
gamma2 ~ dnorm(0, 0.1)

p ~ dbeta(50, 10)

p.val.beta <- step(beta1)

p.val.gamma <- step(gamma1)

}
```

APPENDIX D

Code for Meta Analysis

D.1 MeSH Terms for Literature Search

```
(‘‘surgical procedures, operative’’[MeSH Terms] OR
(‘‘surgical’’[All Fields] AND ‘‘procedures’’[All Fields] AND
‘‘operative’’[All Fields]) OR ‘‘operative surgical procedures’’[All Fields]
OR ‘‘surgical’’[All Fields]) AND ablation[All Fields] AND
(‘‘atrial fibrillation’’[MeSH Terms] OR (‘‘atrial’’[All Fields]
AND ‘‘fibrillation’’[All Fields]) OR ‘‘atrial fibrillation’’[All Fields])
```

D.2 R Code for Meta Analysis

```
library(meta)
meta1 <- metaprop(event, samp_size, studlab=paste(Reference), data=data1)
forest(meta1, rightcols=c("effect", "ci", "w.fixed"), sm="PRAW",
  comb.random=FALSE, leftlabs=c("Author", NA, NA, NA, NA),
  title="Overall Outcome")
funnel(meta1)
```

BIBLIOGRAPHY

(2000), *Diagnostic and statistical manual of mental disorders*, Washington, DC: American Psychiatric Association, revised 4th ed.

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken, New Jersey: Wiley Interscience.

Albage, A., Péterffy, M., and Källner, G. (2011), "The biatrial cryo-maze procedure for treatment of atrial fibrillation: a single center experience," *Scandinavian Cardiovascular Journal*, 45 (2), 112–119.

Anderson, C., Bratcher, T., and Kutran, K. (1994), "Bayesian estimation of population density and visibility," *Texas Journal of Science*, 46, 7–12.

Bagge, L., Blomström, P., Nilsson, L., Einarsson, G. M., Jidéus, L., and Blomström-Lundqvist, C. (2009), "Epicardial off-pump pulmonary vein isolation and vagal denervation improve long-term outcome and quality of life in patients with atrial fibrillation," *Journal of Thoracic and Cardiovascular Surgery*, 137 (5), 1265–1271.

Beavers, D. P. and Stamey, J. D. (2012), "Bayesian sample size determination for binary regression with a misclassified covariate and no gold standard," *Computational Statistics and Data Analysis*, 56, 2574–2582.

Beukema, R., Beukema, W. P., Sie, H. T., Misier, A. R., Delnoy, P. P., and Elvan, A. (2009), "Monitoring of atrial fibrillation burden after surgical ablation: relevancy of end-point criteria after radiofrequency ablation treatment of patients with long atrial fibrillation," *Interactive Cardiovascular and Thoracic Surgery*, 9, 956–959.

Beyer, E., Lee, R., and Lam, B. K. (2009), "Minimally invasive bipolar radiofrequency ablation of lone atrial fibrillation: early multicenter results," *Journal of Thoracic and Cardiovascular Surgery*, 137 (3), 521–526.

Black, M. A. and Craig, B. A. (2002), "Estimating disease prevalence in the absence of a gold standard," *Statistics in Medicine*, 21, 2653–2669.

Borsboom, D. (2005), *Measuring the mind: Conceptual issues in contemporary psychometrics*, Cambridge: Cambridge University.

Branscum, A. J., Johnson, W. O., and Gardner, I. A. (2007), "Sample size calculations for studies designed to evaluate diagnostic test accuracy," *Journal of Agricultural, Biological, and Environmental Statistics*, 12 (1), 112–127.

Brooks, A. G., Stiles, M. K., Laborderie, J., Lau, D. H., Kuklik, P., Shipp, N. J., Hsu, L. F., and Sanders, P. (2010), "Outcomes of long-standing persistent atrial fibrillation ablation: a systematic review," *Heart Rhythm*, 7 (6), 835–846.

Bryant, J. and Day, R. (1995), "Incorporating toxicity considerations into the design of two-stage phase II clinical trials," *Biometrics*, 51 (4), 1372–1383.

Calkins, H., Brugada, J., Packer, D. L., Cappato, R., Chen, S. A., Crijns, H. J., Damiano, R. J., Davies, W., Haines, D. E., Haissaguerre, M., Iesaka, Y., Jackman, W., Jais, P., Kottkamp, H., Kuck, K. H., Lindsay, B. D., Marchlinski, F. E., McCarthy, P. M., Mont, J. L., Morady, F., Nademanee, K., Natale, A., Pappone, C., Prystowsky, E., Raviele, A., Ruskin, J. N., and Shemin, R. J. (2007), "HRS/EHRA/ECAS expert consensus statement on catheter and surgical ablation of atrial fibrillation: recommendations for personnel, policy, procedures and follow-up. A report of the Heart Rhythm Society (HRS) Task Force on catheter and surgical ablation of atrial fibrillation," *Heart Rhythm*, 4, 816–861.

Calkins, H., Reynolds, M. R., Spector, P., Sondhi, M., Xu, Y., Martin, A., Williams, C. J., and Sledge, I. (2009), "Treatment of atrial fibrillation with antiarrhythmic drugs or radiofrequency ablation: two systematic literature reviews and a meta-analysis," *Circulation: Arrhythmia and Electrophysiology*, 2, 349–361.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: a Modern Perspective*, Baco Raton, Florida: Chapman and Hall, 2nd ed.

Chaloner, K. M. and Duncan, G. T. (1983), "Assessment of a beta prior distribution: PM elicitation," *Statistician*, 32, 174–180.

Cheng, D., Stamey, J. D., and Branscum, A. J. (2009), "Bayesian approach to average power calculations for binary regression models with misclassified outcomes," *Statistics in Medicine*, 28, 848–863.

Conaway, M. R. and Petroni, G. R. (1995), "Bivariate sequential designs for phase II trials," *Biometrics*, 51 (2), 656–664.

— (1996), "Designs for phase II trials allowing for a trade-off between response and toxicity," *Biometrics*, 52 (4), 1375–1386.

Cox, J. L., Schuessler, R. B., and Boineau, J. P. (2000), "The development of the Maze procedure for the treatment of atrial fibrillation," *Seminars in Thoracic and Cardiovascular Surgery*, 12 (1), 2–14.

Cui, Y. Q., Li, Y., and Gao, F. (2010), "Video-assisted minimally invasive surgery for lone atrial fibrillation: a clinical report of 81 cases," *Journal of Thoracic and Cardiovascular Surgery*, 139 (2), 326–332.

Davidson, M., Ma, P., Stein, E., Gotta, A., Raza, A., Chitra, R., and Hutchinson, H. (2002), "Comparison of effect on low-density lipoprotein cholesterol and high-density lipoprotein cholesterol with Rosuvastatin versus Atorvastatin in patients with type IIa or IIb hypercholesterolemia," *The American Journal of Cardiology*, 89, 268–275.

Dendukuri, N. and Joseph, L. (2001), "Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests," *Biometrics*, 57, 208–217.

Dendukuri, N., Rahme, E., Belisle, P., and Joseph, L. (2004), "Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard," *Biometrics*, 60, 388–397.

Ederton, Z. J. and Edgerton, J. R. (2008), "Rationale for minimally invasive pulmonary vein isolation and partial autonomic denervation for surgical treatment of atrial fibrillation," *Innovations*, 3, 121–124.

Edgerton, J. R., Brinkman, W. T., Weaver, T., Prince, S. L., Culica, D., Herbert, M. A., and Mack, M. J. (2010), "Pulmonary vein isolation and autonomic denervation for the management of paroxysmal atrial fibrillation by a minimally invasive surgical approach," *Journal of Thoracic and Cardiovascular Surgery*, 140 (4), 823–828.

Edgerton, J. R., Jackman, W. M., and Mack, M. J. (2009a), "A new epicardial lesion set for minimal access left atrial maze: the Dallas Lesion Set," *Annals of Thoracic Surgery*, 88, 1655–1657.

Edgerton, J. R., Jackman, W. M., Mahoney, C., and Mack, M. J. (2009b), "Totally thorascopic surgical ablation of persistent AF and long-standing persistent atrial fibrillation using the Dallas lesion set," *Heart Rhythm*, 6 (12S), S64–S70.

Edgerton, J. R., McClelland, J. H., Duke, D., Gerdisch, M. W., Steinberg, B. M., Bronleewe, S. H., Prince, S. L., Herbert, M. A., Hoffman, S., and Mack, M. J. (2009c), "Minimally invasive surgical ablation of atrial fibrillation: six-month results," *Journal of Thoracic and Cardiovascular Surgery*, 138 (1), 109–114.

Edgerton, Z. J. and Edgerton, J. R. (2009), "History of surgery for atrial fibrillation," *Heart Rhythm*, 6 (12S), S1–S4.

Fader, P. S. and Hardie, B. G. S. (2000), "A note on modeling underreported Poisson counts," *Journal of Applied Statistics*, 8, 953–964.

Fox, J. P. and Glas, C. A. W. (2003), "Bayesian modeling of measurement error in predictor variables using item response theory," *Psychometrika*, 68, 169–191.

Fuster, V., Ryde'n, L. E., Cannom, D. S., Crijns, H. J., Curtis, A. B., Ellenbogen, K. A., Halperin, J. L., Heuzey, J. Y. L., Kay, G. N., Lowe, J. E., Olsson, S. B., Prystowsky, E. N., Tamargo, J. L., Wann, S., Smith, S. C., Jacobs, A. K., Adams, C. D., Anderson, J. L., Antman, E. M., Halperin, J. L., Hunt, S. A., Nishimura, R., Ornato, J. P., Page, R. L., Riegel, B., Priori, S. G., Blanc, J. J., Budai, A., Camm, A. J., Dean, V., Deckers, J. W., Despres, C., Dichstein, K., Lekakis, J., McGregor, K., Metra, M., Morais, J., Osterspreу, A., Tamargo, J. L., Zamorano, J. L., Members, A. T. F., of Cardiology Committee, E. S., and Society, H. R. (2006), "ACC/AHA/ESC 2006 guidlines for the management of patients with atrial fibrillation," *Circulation*, 114, e257–354.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Baco Raton, Florida: Chapman and Hall, 2nd ed.

Gittins, J. and Pezeshk, H. (2000), "A bahvioural Bayes method for determining the size of a clinical trial," *Drug Information Journal*, 34, 355–363.

Go, A. S., Hylek, E. M., Phillips, K. A., Chang, Y., Henault, L. E., Selby, J. V., and Singer, D. E. (2001), "Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study," *Journal of the American Medical Association*, 285, 2370–2375.

Hedeker, D., Gibbons, R. D., and Flay, B. R. (1994), "Random-effects regression models for clustered data with an example from smoking prevention research," *Journal of Consulting and Clinical Psychology*, 62 (4), 757–765.

Jennison, C. and Turnbull, B. W. (1993), "Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints," *Biometrics*, 49 (3), 741–752.

Kazdin, A. E. (2003), *Research Design in Clinical Psychology*, Needham Heights, MA: Allyn & Bacon, 4th ed.

Kikuchi, T. and Gittins, J. (2009), "A behavioral Bayes method to determine the sample size of a clinical trial considering efficacy and safety," *Statistics in Medicine*, 28, 2293–2306.

Kong, M. H., Piccini, J. P., and Bahnson, T. D. (2011), "Efficacy of adjunctive ablation of complex fractionated atrial electrograms and pulmonary vein isolation for the treatment of atrial fibrillation: a meta-analysis of randomized controlled trials," *Europace*, 13, 193–204.

Kron, J., Kasirajan, V., Wood, M. A., Kowalski, M., Han, F. T., and Ellenbogen, K. A. (2009), "Management of recurrent atrial arrhythmias after minimally invasive surgical pulmonary vein isolation and ganglionic plexi ablation for atrial fibrillation," *Heart Rhythm*, 7 (4), 445–451.

Krul, S. P. J., Driessen, A. H., van Boven, W. J., Linnenbank, A. C., Geuzebroek, G. S., Jackman, W. M., Wilde, A. A., de Bakker, J. M., and de Groot, J. R. (2011), "Thoracoscopic video-assisted pulmonary vein antrum isolation, ganglionated plexus ablation, and periprocedural confirmation of ablation lesions: first results of a hybrid surgical-electrophysiological approach for atrial fibrillation," *Circulation: Arrhythmia and Electrophysiology*, 4 (3), 262–270.

Lee, P. (2004), *Bayesian Statistics: an Introduction*, Hodder Arnold, 3rd ed.

Li, W. J., Bai, Y. Y., Zhang, H. Y., Tang, R. B., Miao, C. L., Sang, C. H., Yin, X. D., Dong, J. Z., and Ma, C. S. (2011), "Additional ablation of complex fractionated atrial electrograms after pulmonary vein isolation in patients with atrial fibrillation: a meta-analysis," *Circulation: Arrhythmia and Electrophysiology*, 4, 143–148.

Lord, F. M. and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Mack, M. H. and Holmes, D. R. (2011), "Rational dispersion for the introduction of transcatheter valve therapy," *Journal of American Medical Association*, 306, 2149–2150.

Mack, M. J. (2009), "Current results of minimally invasive surgical ablation for isolated atrial fibrillation," *Heart Rhythm*, 6 (12S), S46–S49.

Meehl, P. E. (1997), *The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions.* In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds), *What if there were no significance tests? (393-425)*, Mahway, NJ: Erlbaum.

Novick, M. R. and Jackson, P. H. (1974), *Statistical methods for educational psychological research*, New York: McGraw-Hill.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics: Bayesian Inference*, London: Arnold.

Oliveira, A., Ramos, E., Lopes, C., and Barros, H. (2009), "Self-reporting weight and height: misclassificaiton effect on the risk estimates for acute myocardial infarction," *European Journal of Public Health*, 19 (5), 548–553.

Olsson, A., Pears, J., McKellar, J., Mizan, J., and Raza, A. (2001), "Effect of Rosuvastatin on low-density lipoprotein cholesterol in patients with hypercholesterolemia," *The American Journal of Cardiology*, 88, 504–508.

Pérez-Stable, E. J., Kaiser, H. J., Marín, G., Marín, B. V., and Benowitz, N. L. (1992), "Misclassification of smoking status by self-reported cigarette consumption," *American Journal of Respiratory and Critical Care Medicine*, 145, 53–57.

Perry, I. and Neuberger, J. (2005), "Immunosuppression: towards a logical approach in liver transplantation," *Clinical and Experimental Immunology*, 139, 2–10.

Press, S. J. (1989), *Bayesian Statistics: Principles, Models, and Applications*, New York, New York: John Wiley & Sons, Inc.

Ramos, F. F. R. (1999), "Underreporting of purchases of port wine," *Journal of Applied Statistics*, 26, 485–494.

Ren, D. and Stone, R. A. (2007), "A Bayesian adjustment for covariate misclassification with correlated binary outcome data," *Journal of Applied Statistics*, 34, 1019–1034.

Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York, New York: Springer Science & Business Media, 2nd ed.

Robert, C. P. (2001), *The Bayesian Choice*, Springer, 2nd ed.

Roy, S., Banjeree, T., and Maiti, T. (2005), "Measurement error model for misclassified binary responses," *Statistics in Medicine*, 24, 269–283.

Saito, Y., Goto, Y., Dane, A., Strutt, K., and Raza, A. (2003), "Randomized dose-response study of Rosuvastatin in Japanese patients with hypercholesterolemia," *Journal of Atherosclerosis and Thrombosis*, 10(6), 329–336.

Scott, A. N., Joseph, L., Bélisle, P., Behr, M. A., and Schwartzman, K. (2008), "Bayesian modelling of tuberculosis clustering from DNA fingerprint data," *Statistics in Medicine*, 27, 140–156.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2011), "Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints," *Journal of Biopharmaceutical Statistics*, 21, 650–658.

Stamey, J. and Katsis, A. (2007), "Sample size determination for comparing two Poisson rates with underreported counts," *Communications in Statistics - Simulation and Computation*, 36, 483–492.

Stamey, J. D., Seaman, J. W., and Young, D. M. (2004), "Bayesian sample size determination for estimating a Poisson rate with underreported data," *Communications in Statistics - Simulation and Computation*, 33(2), 341–354.

Stulak, J. M., Dearani, J. A., Sundt, T. M., Daly, R. C., and Schaff, H. V. (2011), "Ablation of atrial fibrillation: Comparison of catheter-based techniques and the Cox-Maze III operation," *Annals of Thoracic Surgery*, 91, 1882–1889.

Targan, S., Hanauer, S., van Deventer, S., Mayer, L., Present, D., Braakman, T., DeWoody, K., Schaible, T., Rutgeertes, P., and the Crohn's Disease cA2 Study Group F (1997), "A short-term study of chimeric monoclonal antibody cA2 to tumor necrosis factor α for Crohns disease," *The New England Journal of Medicine*, 337, 1029–1035.

Wang, F. and Gelfand, A. E. (2002), "A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models," *Statistical Science*, 17 (2), 193–208.

Weimar, T., Bailey, M. S., Waranabe, Y., Marin, D., Maniar, H. S., Schuessler, R. B., and Damiano, R. J. (2011), "The Cox-maze IV procedure for lone atrial fibrillation: a single center experience in 100 consecutive patients," *Journal of Interventional Cardiac Electrophysiology: an International Journal of Arrhythmias and Pacing*, 31, 47–54.

Whittemore, A. S. and Gong, G. (1991), "Poisson regression with misclassified counts: application to cervical cancer mortality rates," *Applied Statistics*, 40, 81–93.

Winkleman, R. (1996), "Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism," *Empirical Economics*, 21, 575–587.

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001), *Woodcock-Johnson Tests of Cognitive Abilities*, Itasca, IL: Riverside Publishing, 3rd ed.

Yilmaz, A., Greuzebroek, G. S. C., van Putte, B. P., Boersma, L. V., Sonker, U., de Bakker, J. M., and van Boven, W. J. (2010), "Completely thoracoscopic pulmonary vein isolation with ganglionic plexus ablation and left atrial appendage amputation for treatment of atrial fibrillation," *European Journal of Cardio-Thoracic Surgery*, 38 (3), 356–360.