Evaluating the Performance of Hypothesis Testing in Case-Control Studies with Exposure Misclassification, using Frequentist and Bayesian Techniques

by

Mohammad Ehsanul Karim

B.Sc., University of Dhaka, 2004
M.S., University of Dhaka, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2009

Abstract

In epidemiologic studies, measurement error in the exposure variable can have large effects on the power of hypothesis testing for detecting the impact of exposure in the development of a disease. As it distorts the structure of data, more uncertainty is associated with the inferential procedure involving such exposure variables. The underlying theme of this thesis is the adjustment for misclassification in the hypothesis testing procedure. We consider problems involving a correctly measured binary response and a misclassified binary exposure variable in a retrospective case-control scenario. We account for misclassification error via validation data under the assumption of non-differential misclassification. The objective here is to develop a test to check whether the exposure prevalence rates of cases and controls are the same or not, under the frequentist and Bayesian point of view. To evaluate the test developed under the Bayesian approach, we compare that with an equivalent test developed under the frequentist approach. Both these approaches were developed in two different settings: in the presence or absence of validation data, to evaluate whether there is any gain in hypothesis testing for having such validation data. The frequentist approach involves the likelihood ratio test, while the Bayesian test is developed from posterior distribution generated by a mixed MCMC algorithm and a normal prior under realistic assumptions. The comparison between these two approaches is conducted using different simulated scenarios, as well as two real case-control studies having partial validation (internal) data. Different scenarios include settings with varying sensitivity and specificity, sample sizes, exposure prevalence and proportion of unvalidated and validated data. One other scenario that was considered is to evaluate the performance under a fixed budgetary constraint. In the scenarios under consideration, we reach the same conclusion from the two hypothesis testing procedures. The simulation study suggests that the adjusted model (with validation data model) is always better than the unadjusted model (without validation data model). However, exception is possible in the fixed budget scenario.

Table of Contents

Abstract ii

Table of Contents iii

List of Tables v

List of Figures vi

Selected Notations viii

Acknowledgements ix

1 Prelude 1
- 1.1 Introduction 1
- 1.2 The Impact of Misclassification 2
- 1.3 Suggested Correction of Misclassification in the Literature . 3
- 1.4 Settings of the Problem under Investigation 4
- 1.5 Basic Terminologies used to Evaluate Misclassification and Measures of Association 7
- 1.6 Existing Literature on Misclassification 10
- 1.7 Motivation and Outline of the Current Work 12

2 Frequentist Adjustment 13
- 2.1 Introduction 13
- 2.2 Likelihood Functions 13
  - 2.2.1 Without Validation Data 15
  - 2.2.2 With Validation Data 15
- 2.3 Variance Estimates 17
- 2.4 Likelihood-ratio Tests 17

3 Bayesian Adjustment 19
- 3.1 Introduction 19
  - 3.1.1 Bayes' Theorem 19

3.2 MCMC Algorithms 19
3.2.1 Metropolis-Hastings Algorithm 21
3.2.2 Gibbs Algorithm 22
3.2.3 Mixed Algorithm 23
3.3 MCMC Diagnostics 29
3.3.1 Conventional Graphical Diagnosis 31
3.3.2 Gelman-Rubin Method for Monitoring Convergence . 31

4 Simulation Results 33
4.1 Data Generation 33
4.2 Scenario Settings Under Frequentist Adjustment 33
4.2.1 Under Different Values of Sensitivity and Specificity . 35
4.2.2 Under Different Sample Sizes 36
4.2.3 Under Different Exposure Prevalence Rates 36
4.2.4 Under Different Proportion of Validation and Main Part of the Data 39
4.2.5 Comparison under Budgetary Constraint 39
4.3 Scenario Settings Under Bayesian Adjustment 48
4.3.1 Under Different Values of Sensitivity and Specificity . 49
4.3.2 Under Different Sample Sizes 49
4.3.3 Under Different Exposure Prevalence Rates 49
4.3.4 Under Different Proportion of the Validation Data . . 49
4.3.5 Diagnostics 54

5 Application in Epidemiological Studies 64
5.1 Introduction 64
5.2 Study of Sudden Infant Death Syndrome (SIDS) 64
5.3 Cervical Cancer and Herpes Simplex Virus Study 68

6 Conclusions and Further Research 75
6.1 Overall Conclusions 75
6.2 Further Research and Recommendations 77

Bibliography 79

List of Tables

1.1 Structure for main (unvalidated) part of the data 6
1.2 Structure for validation part of the data 7
1.3 Relationships among the basic terminologies in a 2×2 table . 7

4.1 Scenarios under consideration 35
4.2 Scenarios under constant cost = \$2400, assuming that collecting validated data costs three (3) times as much as collecting unvalidated (main) data 43
4.3 Scenarios under constant cost = \$2400, assuming that collecting validated data costs five (5) times as much as collecting unvalidated (main) data 44
4.4 Scenarios under constant cost = \$2400, assuming that collecting validated data costs ten (10) times as much as collecting unvalidated (main) data 45

5.1 Data from the study of sudden infant death syndrome (SIDS) and antibiotic prescription 65
5.2 Frequentist Estimates of the model parameters in the SIDS study 65
5.3 Bayesian Estimates of the model parameters in the SIDS study 68
5.4 Data from Herpes Simplex Virus-2 study 68
5.5 Frequentist Estimates of the model parameters in the HSV-2 study 70
5.6 Bayesian Estimates of the model parameters in the HSV-2 study 71

List of Figures

4.1 Power curves under different sensitivity and specificity values: 0.6, 0.7, 0.8 and 0.9 respectively 37
4.2 Power curves under different sample sizes: 400, 600, 1000 and 2000 respectively (validation sub-sample size is fixed at 200 in each situation) 38
4.3 Power curves under different Exposure Prevalences: 0.25, 0.30, 0.35 and 0.4 respectively 40
4.4 Power curves under smaller Exposure Prevalences: 0.005, 0.01, 0.05 and 0.10 respectively 41
4.5 Power curves under different proportions of validation part and main part of the data: (1:9, 1:3, 1:1 and 3:1) respectively 42
4.6 Power curves under fixed amount of cost = $2400 assuming that collecting validated data costs three (3) times as much as collecting unvalidated (main) data 44
4.7 Power curves under fixed amount of cost = $2400 assuming that collecting validated data costs five (5) times as much as collecting unvalidated (main) data 45
4.8 Power curves under fixed amount of cost = $2400 assuming that collecting validated data costs ten (10) times as much as collecting unvalidated (main) data 46
4.9 Bayesian analysis results for different sensitivity and specificity values (.6, .7, .8, .9): Proportion of credible intervals excluding null value 50
4.10 Bayesian analysis results for different sample sizes (400, 600, 1000, 2000): Proportion of credible intervals excluding null value 51
4.11 Bayesian analysis results for different exposure prevalence (.25, .3, .35, .4): Proportion of credible intervals excluding null value 52

4.12 Bayesian analysis results for different ratio of data splits (1:9, 1:3, 1:1, 3:1 respectively for validation and main part): Proportion of credible intervals excluding null value 53
4.13 Diagnosis of convergence of Bayesian analysis results: Trace Plots for r_0 in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset . . . 55
4.14 Diagnosis of convergence of Bayesian analysis results: Trace Plots for r_1 in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset . . . 56
4.15 Diagnosis of convergence of Bayesian analysis results: Trace Plots for SN in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset . . . 57
4.16 Diagnosis of convergence of Bayesian analysis results: Trace Plots for SP in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset . . . 58
4.17 Sequence of the mean of posterior for r_0 for the four Markov Chain Monte Carlo Chains for 10,000 iterations 59
4.18 Sequence of the mean of posterior for r_1 for the four Markov Chain Monte Carlo Chains for 10,000 iterations 60
4.19 Sequence of the mean of posterior for SN for the four Markov Chain Monte Carlo Chains for 10,000 iterations 61
4.20 Sequence of the mean of posterior for SP for the four Markov Chain Monte Carlo Chains for 10,000 iterations 62
4.21 Sequence of the Gelman-Rubin $\hat{R}$ for the four Markov Chain Monte Carlo Chains for 10,000 iterations 63

5.1 MCMC for the with and without validation data model parameters in the SIDS study 67
5.2 Prior and Posterior Distributions of all the Parameters under Consideration in the SIDS study 69
5.3 MCMC for the with and without validation data model parameters in the HSV-2 study 72
5.4 Prior and Posterior Distributions of all the Parameters under Consideration in the HSV-2 study 74

Selected Notations

Y	=	Outcome variable of a study
V	=	Categorical explanatory variable
V^*	=	Recorded surrogate variable which is collected instead of V for practical reasons
$n_{..}$	=	Observed part of the data
$u_{..}$	=	Unobserved part of the data
r	=	Exposure prevalence
θ	=	Apparent exposure prevalence
SN	=	Sensitivity
SP	=	Specificity
Ψ	=	Odds-ratio
PPV	=	Positive predictive value
NPV	=	Negative predictive value
$L(.)$	=	Likelihood function
$f(.)$	=	density function
Π	=	Logit transformation of exposure prevalence
Γ	=	Logit transformation of sensitivity
Υ	=	Logit transformation of specificity
Θ	=	Logit transformation of apparent exposure prevalence
i, j, K, t	=	Index

Acknowledgements

I would like to thank everyone of the Department of Statistics, from faculty to staff and fellow graduate students, for making my M.Sc. program such an enriching and pleasant experience. In particular, it gives me a great pleasure to express my sincere gratitude and deepest appreciation to my supervisor Professor Paul Gustafson. His unique way of mentoring, valuable suggestions regarding my research, technical issues of research, hints for programming calculations, careful draft revision, constant inspiration and words of wisdom helped me a great deal to complete this dissertation in time. It truly has been an honor and a privilege to work with him.

Nonetheless, I would like to express my sincere gratitude and appreciation to Professors John Petkau, Jiahua Chen, Matías Salibián-Barrera, Ruben H. Zamar, Arnaud Doucet, Lang Wu and Michael Schulzer for their excellent teaching and mentorship. Again, I would like to thank Professor John Petkau for kindly agreeing to be the second reader of my thesis, and for his numerous suggestions regarding clarity and consistency in my writing.

On a personal note, I am eternally grateful to my parents who have provided me with abundance of opportunities and freedom all my life, and to my wife Suborna for her love.

Vancouver,
Canada
August 27, 2009

Mohammad Ehsanul Karim
ehsan@stat.ubc.ca

To my parents, and my wife Suborna

Chapter 1

Prelude

1.1 Introduction

A health outcome, often simply presence or absence of disease, is usually the central issue of an epidemiological inquiry, whereas the exposure is a related factor that is possibly involved in development of disease. The scope of an exposure assessment is much broader in the sense that it can originate from various sources such as some physiological characteristic, psychological characteristic, genetic factor, social or environmental element or genetic attribute. We can use some biological test or even self-reported survey instrument to assess exposure status. Intuition suggests that, whatever tool we use to evaluate that exposure status, there is always a possibility of having mismeasurement. We can have a gold standard method of exposure status evaluation with a well-set definition of superior or ideal exposure assessment. However, since such superior assessment may not be possible to implement on the whole study sample for various practical reasons, such as available resources or ethical considerations, an operational method of assessment has to be settled upon so that we can use that method on the entire sample. This operational definition is basically an indirect measure of the exposure of ultimate interest. The methodologies of assessing disease and evaluating exposure are quite different from one another. Therefore, the mechanisms by which measurement errors will occur from these two sources are very different.

Evaluating the affect of the exposure to a given risk factor in the development of a disease or infection is usually the goal in epidemiological studies. While making the causal association between an outcome variable that defines the disease and the exposure variable(s), it is crucial that both are recorded without error. However, due to restriction of resources, often such quantification by any association measure is hindered by the lack of preciseness of the measures of relevant exposures which are collected using the operational definition. When there exist any sources of error, it is possible that the researcher's interpretations or findings of causal inference

have alternative explanations. A rich literature suggests that this has long been identified as a problem and there has been considerable interest in this problem. Most applied work still ignores this issue and suffers detrimental effects. In the current work, this issue is acknowledged and addressed.

This problem is relevant as much to continuous as to categorical measures – although the terminology differs slightly. 'Measurement error' is the terminology used when the predictor variable under consideration is continuous in nature. On average, the closer the true explanatory variable value and the measured value from the surrogate variable (error-corrupted variable) are, the less measurement error exits. If the predictor variable is categorical instead (with two or more categories), we call it a problem of 'misclassification'. In this case, the probabilities of classifying a subject into the correct category are considered. The impact of both of the mismeasurement cases are somewhat similar, although the expressions and terminologies to evaluate them are quite different.

For the sake of clarity, let us define some notation: the goal of the study is to explain the relationship between the outcome variable (Y) and exposure variable. In the current work, we will restrict ourselves to a binary exposure variable and denote it as V. That is, we will only consider the situation whether a subject is either exposed or not. However, for practical reasons such as cost, time factors or unavailability of a gold standard, V might not be measured precisely or directly. Therefore, a cruder classification method is applied and a corresponding surrogate variable V^* is recorded instead. This is mostly the case when the exposure status is unobservable or cannot be measured precisely within reasonable cost. Nevertheless, although plugging-in a surrogate variable by using an imprecise but cheap classification tool might seem a very intuitive solution, this is are not without consequences. The phenomenon of such error on the measure of association is sometimes referred to as information bias. A question of accuracy of the estimate of the measure of association between disease and exposure arises, and hence we need to evaluate the impact of such replacements.

1.2 The Impact of Misclassification

Mismeasurement in the explanatory variables, when ignored, can have detrimental effects on statistical analysis such as: making the estimates of the

parameters biased in the model under investigation, reducing the discriminative power, and masking various features of the data [*Carroll et al.*, 2006].

In the case of obtaining an estimate of a measure of association, misclassification presents a serious problem. Naive analysis that just substitutes the apparent exposure status for the unobserved true exposure status can produce highly biased estimates. When misclassification probabilities are equal for the two compared groups (exposed and unexposed), the estimates of measures of association such as relative risk, odds ratio, are biased toward the null value [*Copeland et al.*, 1977].

However, the effect of misclassification error on hypothesis testing procedures might not be as detrimental as that on estimation, as mentioned in *Bross* [1954]. In this paper, it is argued that, if similar misclassification prevails in both exposed and unexposed populations, then the validity of the test of finding whether two proportions, that is, the exposure prevalences are different or not, is not affected. However, this does not come without a price - and the price is the power of the test, which is reduced in the presence of misclassification. Usually the extent of loss depends on the amount of misclassification.

1.3 Suggested Correction of Misclassification in the Literature

Fortunately, reasonable estimates of measures of association are still attainable, even though the exposure variable under consideration is corrupted. For that, the researchers must have some knowledge about the nature of error to be able to correct or account for it. Identifiability becomes an issue for the likelihoods - if we have no clue regarding the extent of misclassification [*Walter and Irwig*, 1988]. A number of methods for the correction of measurement error have been developed throughout the years, both in design and analysis stages. Methodologies in the design stage include replicated measurements, validation studies, etc. In the validation study, the validated sub-sample is derived randomly from the same population under investigation (either internal or external to those included in the primary sample) and a superior method of exposure assessment is implemented on each subject in the sub-sample. All these methods have their own pros and cons. Taking into account such information, correction for misclassification or measurement error can be performed either in frequentist or Bayesian

ways. We will discuss this topic in the current work from the 'test of hypothesis' point of view.

It is worth mentioning that we will be using internal validation sub-samples throughout this work. Although external validation sample sometimes helps generalizing the results to larger extent, it suffers from various other limitations as well, especially in the situations when cause is dependent on many factors, not only on the predictor variable under investigation, which is a very common circumstance in the disease-exposure relationships. Also, in terms of cost, internal validation sub-samples are cheaper than external validation samples since inferior method of exposure assessment is already applied on the subjects of internal validation sub-samples.

1.4 Settings of the Problem under Investigation

Let Y be the outcome of interest:

$$Y = \begin{cases} 1 & \text{for diseased subjects} \\ 0 & \text{for disease free subjects.} \end{cases}$$

To keep the problem simple, it is assumed that the outcome variable Y is measurement error free. That is, we will deal with exposure misclassification, not disease misclassification.

The simplest setting in misclassification is a binary exposure variable, which is frequently the case in epidemiological studies. The binary variable V is used to denote the true exposure status:

$$V = \begin{cases} 1 & \text{for truly exposed} \\ 0 & \text{for truly unexposed.} \end{cases}$$

V^* is a surrogate variable that denotes the exposure status observed by some instrument or measurement that is subject to a certain amount of error:

$$V^* = \begin{cases} 1 & \text{for apparently exposed} \\ 0 & \text{for apparently unexposed.} \end{cases}$$

Here the exposure variable V is considered to be replaced by the surrogate variable V^* with considerable measurement error. It is also assumed that such exposure measurements are independent of other errors.

To obtain information about the degree of mismeasurement, a validation sub-sample is necessary, where complete information is available about true exposure status (V), along with surrogate variable V^* status (through an imperfect assessment on exposure). This is a small fraction of the main sample, where only the surrogate variable V^* status is available. Throughout this work, we used various compositions of data by varying this fraction. We will discuss this further in Chapter 4.

Although it is known that prospective study data are usually preferable study data, researchers have to make certain trade-offs due to feasibility. Retrospective designs are more popular because the secondary data sources are usually much cheaper. However, (unmatched) retrospective case-control studies are more subject to errors of measurement or misclassification, which often leads to invalid results. Therefore, we consider a retrospective case-control scenario, where n_1 subjects are sampled from the diseased population (cases), and n_0 subjects are sampled from the disease free population (controls).

To make valid causal inference from a retrospective study, a number of assumptions need to be appreciated. Consideration of the type or pattern of measurement error is very crucial in evaluating its likely impact on a measure of association. Researchers should be able to distinguish the consequences of different patters of misclassification: such as differential and nondifferential misclassification - which are based on whether the pattern of error in exposure assessment varies with respect to disease status. Misclassification probabilities of exposure vary with respect to disease status in case of differential misclassification. Errors arising due to recall bias and perception are common sources for misclassification probabilities being different in relation to disease status. The presence of disease may have great influence on how subjects interpret or report about the exposure status. In this case, the conditional distribution of the surrogate exposure variable (or measurement by 'imperfect' exposure assessment method), given the true exposure variable and outcome variable, that is, $V^*|V, Y$, depends on Y. This is the case for many realistic situations. However, to simplify the problem, we sometimes assume that the conditional distribution of $V^*|V, Y$ does not depend on Y, that is, misclassification probabilities are invariant with respect to disease status (all cases and controls have the same probability of being misclassified). This is the definition of nondifferential misclassification. Throughout the current work, we will maintain the assumption of nondifferential misclassification, and the conclusions are valid under this particular

assumption. The reason for this assumption is basically due to some of the simple features that are established in literature, such as "bias toward the null" in absence of other forms of error for a dichotomous exposure variable and its simplicity compared to relatively unpredictable effects of differential misclassification. Researchers usually go through more sophisticated designs like blinding (of exposure assessment with respect to the disease outcome) or some of its advanced variants to attempt to ensure that the nondifferential assumption holds.

The notation we will use for the unvalidated sub-sample and validation sub-sample data structures under consideration is given in Table 1.1 and Table 1.2 respectively. The unvalidated and validation sub-samples are separate parts of the entire data. The validation sub-sample is the part of the data where we implemented both the inferior and superior methods of exposure assessment. On the other hand, the unvalidated sub-sample is the part on which we used only the inferior method of exposure assessment, excluding those subjects who were randomly selected for the validation sub-sample. For convenience, we will use the phrase 'unvalidated sub-sample data' and 'main data' interchangeably from now on. In each of these tables, the n_{ij}'s are observed, where $i = 0, 1$, $j = 1, 2, 3, 4, 5, 6$, but in Table 1.1, the u_{ij}'s are unobserved. Although the marginal totals of V^* are observable, we do not have direct information on how those subjects are classified with respect to V. The total number of subjects in the case group is $n_{11} + n_{12} + n_{13} + n_{14} + n_{15} + n_{16} = n_1$ and similarly, the total number of subjects in the control group is $n_{01} + n_{02} + n_{03} + n_{04} + n_{05} + n_{06} = n_0$.

Table 1.1: Structure for main (unvalidated) part of the data

Y	$Y = 1$		$Y = 0$	
V / V^*	$V^* = 1$	$V^* = 0$	$V^* = 1$	$V^* = 0$
$V = 1$	u_{11}	u_{12}	u_{01}	u_{02}
$V = 0$	u_{13}	u_{14}	u_{03}	u_{04}
Total	$u_{11} + u_{13}$ $= n_{15}$	$u_{12} + u_{14}$ $= n_{16}$	$u_{01} + u_{03}$ $= n_{05}$	$u_{02} + u_{04}$ $= n_{06}$

Table 1.2: Structure for validation part of the data

Y	$Y=1$		$Y=0$	
V / V^*	$V^*=1$	$V^*=0$	$V^*=1$	$V^*=0$
$V=1$	n_{11}	n_{12}	n_{01}	n_{02}
$V=0$	n_{13}	n_{14}	n_{03}	n_{04}
Total	$n_{11}+n_{13}$	$n_{12}+n_{14}$	$n_{01}+n_{03}$	$n_{02}+n_{04}$

Table 1.3: Relationships among the basic terminologies in a 2×2 table

		Test Condition	
		Exposed	Unexposed
True Condition	Exposed	True Positive	False Negative
	Unexposed	False Positive	True Negative

1.5 Basic Terminologies used to Evaluate Misclassification and Measures of Association

Let us denote the true exposure prevalence as:

$$r_i = P(V = 1|Y = i),$$

where $i = 0, 1$ for control and case respectively. As V in this case is unobserved, the apparent exposure prevalence is defined as

$$\theta_i = P(V^* = 1|Y = i).$$

Sensitivity and specificity are commonly used statistical measures of the performance of a binary classification test. In the current context, sensitivity (SN_i) measures the proportion of actual exposed people which are correctly identified as such. Specificity (SP_i) measures the proportion of unexposed people which are correctly identified. Thus, by definition,

$$SN_i = P(V^* = 1|V = 1, Y = i)$$
$$SP_i = P(V^* = 0|V = 0, Y = i)$$

Notice that we are characterizing misclassification in terms of classification probabilities. Therefore, SN_i and SP_i range between 0 and 1, and the extent to which these are less than 1 indicates the intensity of the misclassification problem.

When the conditional distribution of $V^*|V,Y$ does not depend on Y (i.e., nondifferential misclassification condition), then we get $SN_0 = SN_1 = SN$ and $SP_0 = SP_1 = SP$. The apparent exposure prevalence θ_i can be expressed in terms of r_i, SN_i, SP_i:

$$\begin{aligned}
\theta_i &= P(V^* = 1|Y = i) \\
&= \sum_{k=0}^{1} P(V^* = 1, V = k|Y = i) \\
&= \sum_{k=0}^{1} P(V^* = 1|V = k, Y = i)P(V = k|Y = i) \\
&= P(V^* = 1|V = 1, Y = i)P(V = 1|Y = i) + \\
&\quad P(V^* = 1|V = 0, Y = i)P(V = 0|Y = i) \\
&= SN_i\, r_i + (1 - SP_i)(1 - r_i) \qquad (1.1) \\
&= SN\, r_i + (1 - SP)(1 - r_i),
\end{aligned}$$

denoting common sensitivity by SN and common specificity by SP, under the assumption of nondifferential classification. Simple algebraic manipulation from Equation (1.1) shows that (r_i, SN_i, SP_i) and $(1 - r_i, 1 - SN_i, 1 - SP_i)$ leads to same θ_i.

From Youden's Index [*Youden*, 1950], we know that if the sensitivity and specificity are such that $SN + SP - 1 < 0$, then the test is misleading. $SN + SP = 1$ would mean that the test is no more useful than a coin-flip guess. That is, the test has no discriminative power on the exposure group, and reports same proportion of positive tests for both exposed and unexposed groups. Therefore, a common assumption is $SN + SP > 1$.

In our scenario, where both the response Y and the exposure variable V are binary, the odds-ratio is defined as:

$$\begin{aligned}
\Psi &= \frac{P(V = 1|Y = 1)/P(V = 0|Y = 1)}{P(V = 1|Y = 0)/P(V = 0|Y = 0)} \\
&= \frac{r_1/(1 - r_1)}{r_0/(1 - r_0)}, \qquad (1.2)
\end{aligned}$$

which is a common measure of association between disease and exposure status for retrospective case-control studies. However, if exposure variable V is subject to misclassification error, an intuitive substitute is:

$$\Psi^* = \frac{\theta_1/(1 - \theta_1)}{\theta_0/(1 - \theta_0)}. \qquad (1.3)$$

Thus, the attenuation factor is defined as:

$$AF = \frac{\Psi^*}{\Psi},$$

which gives us an idea of the magnitude of bias introduced by misclassification.

An alternative formulation for expressing degree of misclassification requires us to use the Positive Predictive Value (PPV_i) and the Negative Predictive Value (NPV_i). Positive Predictive Value (PPV_i) is the proportion of subjects with a positive test result from the inferior method of exposure assessment, who actually is exposed, determined by superior method of exposure assessment. Similarly, Negative Predictive Value (NPV_i) is the proportion of subjects with a negative test result from the inferior method of exposure assessment, who actually is unexposed, as indicated by superior method of exposure assessment. These two quantities can be calculated from a 2×2 table. By implementing Bayes' Rule as discussed in Equation (3.1), the relationships of (r_i, SN_i, SP_i) with PPV_i and NPV_i are derived as follows:

$$\begin{aligned} PPV_i &= P(V = 1|V^* = 1, Y = i) \\ &= \frac{P(V^* = 1|V = 1, Y = i)P(V = 1|Y = i)}{P(V^* = 1|V = 1, Y = i)P(V = 1|Y = i) + P(V^* = 1|V = 0, Y = i)P(V = 0|Y = i)} \\ &= \frac{SN_i r_i}{SN_i r_i + (1 - SP_i)(1 - r_i)}. \end{aligned} \tag{1.4}$$

$$\begin{aligned} NPV_i &= P(V = 0|V^* = 0, Y = i) \\ &= \frac{P(V^* = 0|V = 0, Y = i)P(V = 0|Y = i)}{P(V^* = 0|V = 0, Y = i)P(V = 0|Y = i) + P(V^* = 1|V = 0, Y = i)P(V = 0|Y = i)} \\ &= \frac{SP_i(1 - r_i)}{SP_i(1 - r_i) + (1 - SN_i)r_i}. \end{aligned} \tag{1.5}$$

However, unlike the implication of nondifferential misclassification with respect to sensitivity SN and specificity SP, PPV_0 does not have to be equal to PPV_1, nor does NPV_0 has to be equal to NPV_1 under nondifferential misclassification.

1.6 Existing Literature on Misclassification

Nondifferentiality is a recurring assumption in the epidemiologic literature due to some of its interesting results. *Bross* [1954] discussed the difficulties of inferences on a single proportion or the difference between two proportions from a 2×2 classification table in the presence of misclassification. He indicated the distortion of estimation and the power reduction in hypothesis testing. He justified his statements under the assumption of nondifferential misclassification. *Newell* [1962] further substantiated the fact that nondifferential misclassification errors will always tend to produce results biased towards the null (that is, the difference between the two rates will shrink while applying inferential procedures on the data with nondifferential misclassification). Also, *Gullen et al.* [1968] suggested that under broad assumptions, classification error never results in the apparent difference being larger than the real difference in rates. *Dosemeci et al.* [1990] and *Diamond and Lilienfeld* [1962a, b] showed with some numerical examples that exceptions are possible and that nondifferentiality is not always tenable. *Keys and Kihlberg* [1963] tried to identify the reasons of such unusual deviation. To implement this result, the measurement error has to be independent of all other errors. A few good reviews of such unusual phenomenon are available in the literature, such as *Thomas* [1995] and *Jurek et al.* [2005].

Rothman et al. [2008] discussed misuse of the "bias toward the null" result, mostly when the assumptions for this result are not met. Even if the assumptions are met, it is not necessarily true for hypothesis testing: p-values need not have upward bias as reported by *Greenland and Gustafson* [2006].

The situation gets even more complicated for more than two categories, i.e., when exposure is polytomous. [*Gladen and Rogan*, 1979] provided expressions for bias under nondifferential assumption. Early literature on the impact of misclassification includes *Koch* [1969] and *Goldberg* [1975]. Most of these describe the effect on association measures obtained from a 2×2 exposure-disease classification table. *Goldberg* [1972] discusses the issue with regard to hypothesis testing.

Historically, the development of adjustments for mismeasurement were mostly under the nondifferentiality assumption. *Copeland et al.* [1977] suggested extension of the "bias toward the null" result to ratio effect measures of association, such as the risk ratio and odds ratio and derived adjust-

ment formulas to correct for misclassification given the nondifferential assumption. *Barron* [1977] suggested a matrix method for such adjustment. *Greenland* [1980] further extended the adjustment to difference effect measures and also considered the possibility of misclassification of confounders. *Greenland* [1988b] discussed the basic methods for constructing variance estimators for the various parameters after adjusting for misclassification. *Marshall* [1990] proposed inverse matrix methods by reparameterizing the misclassification problem. *Morrissey and Spiegelman* [1999] discussed both the matrix and inverse matrix methods under various circumstances. *Lyles* [2002] reparameterized the likelihood of the problem and suggested a relatively more convenient solution to the problem which does not require numerical optimization. If all the parameters are unknown, nonidentifiability makes the inference impossible. A reasonable estimate of the misclassification probabilities is required to carry on the inference. Adjustments for misclassification using replicated samples are provided by *Walter and Irwig* [1988]. *Greenland* [1988a] provided formulas for adjustment when a validation sample is present. More recent works include *Greenland and Gustafson* [2006], *Greenland* [2008] and *Marshall* [1997]. *Marshall* [1989] pointed out that the estimates of measures of association that adjust for misclassification are very sensitive to the estimates of misclassification probabilities and even small discrepancies with actual probabilities can lead to misadjustment.

Recent developments in the rapidly advancing field of computing made it possible to use the numerical approaches and simulation techniques to solve these problems in a more elegant way. The problems of mismeasurement were explored from a Bayesian context in *Rahme et al.* [2000], *Joseph et al.* [1995] and *Prescott and Garthwaite* [2002]. *Gustafson et al.* [2001] checked the point made by *Marshall* [1989] and suggested a Bayesian solution of the problem by incorporating some uncertainty about the misclassification probabilities by means of having a prior distribution of those parameters instead of a particular guess. *Gustafson and Greenland* [2006] showed that implementing such prior may provide narrower interval estimates of measure of association. *Chu* [2005] incorporated such uncertainty or randomness by implementing various prior distributions on the prevalence and misclassification probabilities and assessed the Bayesian adjustment of estimates of various parameters of misclassified data under various assumptions when validation data is available. Estimates obtained from the Bayesian approach is then compared with estimates from previously developed methods such as the maximum likelihood estimates [*Lyles*, 2002] and SIMEX (simulation extrapolation method).

A general overview of the methods for misclassified categorical data and some extensions to higher dimensions are provided in *Willett* [1989] and *Chen* [1989]. Overall general discussion of these issues and the ways to combat such problems are documented in chapters 3 and 5 of *Gustafson* [2004].

1.7 Motivation and Outline of the Current Work

Although comparisons between the frequentist method with specific estimates of parameters (misclassification probabilities and prevalences) and the Bayesian method with prior distributions on parameters (to incorporate uncertainty) provided in the literature, such comparisons have not yet been made for hypothesis testing. In this thesis, we will assess the impact of misclassification of dichotomous exposure on hypothesis testing for two settings - without considering validation data and its counterpart after adjustments using the estimates from validation data - under the nondifferential misclassification assumption. The Bayesian adjustments for hypothesis testing will be compared with standard frequentist methods.

In Chapter 1, we have discussed the historical developments, basic definitions and terminologies for misclassification error. The motivations for correction and some methods of adjusting for such errors are also discussed. The problem under investigation is specified. In Chapter 2 and 3, we will explain the models and methodologies of hypothesis testing in the presence of misclassification error from the frequentist and Bayesian points of view respectively. In chapter 4 we will show the simulation results under a set of scenarios and compare the classical and Bayesian methods. We use some real epidemiological datasets to implement these methods in Chapter 5 and conclude with general findings and further recommendations for future researches in Chapter 6.

Chapter 2

Frequentist Adjustment

2.1 Introduction

Maximum likelihood estimation (MLE) is a popular method used for fitting a statistical model to data. Pioneered by various statisticians including R. A. Fisher at the beginning of the last century, it has widespread applications in various fields. If the sample observations are available, this estimation procedure searches over various possible population characteristics and eventually obtains the most likely value as the estimate of that population characteristic. Having drawn a sample of n values x_1, x_2, ..., x_n from a distribution where ϕ is the parameter of interest, we form $L(\phi) = f(x_1, x_2, \ldots, x_n)$. The method of maximum likelihood estimates ϕ by finding the value of ϕ that maximizes $L(\phi)$ or, more commonly, the logarithmic transformed version of it. The solution can be found numerically using various optimization algorithms. The popular alternatives to this estimation procedure are least squares procedure and method of moments. However, those estimates are not very efficient in various circumstances, whereas maximum likelihood estimates possess various desirable features such as consistency and asymptotically efficiency, if solution exists. The maximum likelihood estimation procedure can also be used on non-random samples, if certain adjustments are made, such as conditioning on the clusters or correlated groups, etc.

2.2 Likelihood Functions

Previously in §1.2, we discussed the impact of misclassification. The estimates of (θ_0, θ_1) obtained from the entire sample will be biased toward the null, under certain conditions. As described in §1.3, there are various methods suggested in the literature for adjusting the consequences of misclassification. We will use the method that uses a validation sub-sample. By using validation data, we can have an estimate of r_0, r_1, SN and SP. Therefore, given the observed data, under nondifferential misclassification, we can consider (r_0, r_1) or (θ_0, θ_1) as the unknown parameters in the statis-

tical models. As mentioned in §1.5, the true exposure prevalence is defined as $r_i = P(V = 1|Y = i)$, whereas, the apparent exposure prevalence is expressed as $\theta_i = P(V^* = 1|Y = i)$. From Equation (1.1), we can see that θ_i can be expressed as a linear function of (r_i, SN, SP). For the same dataset - when r_i, SN and SP remains fixed, we can use $H_0 : \theta_0 = \theta_1$ for without validation data settings (by ignoring all the true exposure status V, but using all the apparent exposure status V^* obtainable from the entire sample) and equivalently, $H_0 : r_0 = r_1$ for with validation data settings (by incorporating the true exposure status V from the validation sub-sample and the apparent exposure status V^* obtainable from the entire sample). Since both of these hypotheses are applied on the same dataset, the total number of subjects under consideration are the same in each test. The stated hypotheses are simply variants of the following hypotheses respectively: $H_0 : \Psi = 1$ and $H_0 : \Psi^* = 1$.

The notable distinction between these two models is that r_i can take any value between $(0, 1)$, whereas θ_i can take values between $\min(SN, 1 - SP)$ and $\max(SN, 1 - SP)$. We will discuss the likelihoods and the solution methods in the following subsections.

One important point is worth mentioning: even in absence of validation data (that is, when true r_0, r_1 , SN and SP are not estimable), due to the equivalence of hypotheses mentioned above, we can test $H_0 : \theta_0 = \theta_1 = \theta$ and we can conclude the same about $H_0 : r_0 = r_1 = r$. However, when validation data is not present, such equivalence is not true for estimation purposes, because when SN and SP are unknown, the relationship between (θ_0, θ_1) and (r_0, r_1) is not known respectively (see Equation 1.1). Therefore, when a validation sub-sample is not available, we can not estimate (r_0, r_1), but from the entire sample we can estimate (θ_0, θ_1).

2.2.1 Without Validation Data

A standard way to express the likelihood in terms of the parameters (θ_0, θ_1) for problems consisting of misclassified data without validation part is:

$$\begin{aligned} & L(\theta_0, \theta_1 | V^*, Y) \\ & \propto \; P(V^* = 1|Y = 0)^{(n_{01}+n_{03}+n_{05})} \times P(V^* = 0|Y = 0)^{(n_{02}+n_{04}+n_{06})} \times \\ & \quad P(V^* = 1|Y = 1)^{(n_{11}+n_{13}+n_{15})} \times P(V^* = 0|Y = 1)^{(n_{12}+n_{14}+n_{16})} \\ & = \; \theta_0^{(n_{01}+n_{03}+n_{05})} \times \{1 - \theta_0\}^{(n_{02}+n_{04}+n_{06})} \times \\ & \quad \theta_1^{(n_{11}+n_{13}+n_{15})} \times \{1 - \theta_1\}^{(n_{12}+n_{14}+n_{16})}. \end{aligned} \tag{2.1}$$

The maximum likelihood estimates of θ_0, θ_1 respectively are given by:

$$\begin{aligned} \hat{\theta}_0 &= \frac{n_{01} + n_{03} + n_{05}}{n_{01} + n_{02} + n_{03} + n_{04} + n_{05} + n_{06}}, \\ \hat{\theta}_1 &= \frac{n_{11} + n_{13} + n_{15}}{n_{11} + n_{12} + n_{13} + n_{14} + n_{15} + n_{16}}. \end{aligned}$$

Under the null hypothesis $H_0 : \theta_0 = \theta_1 = \theta$, the maximum likelihood estimate is given by -

$$\hat{\theta} = \frac{n_{01} + n_{03} + n_{05} + n_{11} + n_{13} + n_{15}}{n_{01} + n_{02} + n_{03} + n_{04} + n_{05} + n_{06} + n_{11} + n_{12} + n_{13} + n_{14} + n_{15} + n_{16}}.$$

2.2.2 With Validation Data

A standard way to express the likelihood in terms of the parameters (r_0, r_1, SN, SP) for problems consisting of misclassified data with validation part

is provided in Equation (2.2) under nondifferential misclassification:

$$
\begin{aligned}
& L(r_0, r_1, SN, SP | V^*, V, Y) \\
\propto\ & \{P(V=1|Y=0)P(V^*=0|V=1,Y=0)\}^{n_{01}} \times \\
& \{P(V=1|Y=0)P(V^*=1|V=1,Y=0)\}^{n_{02}} \times \\
& \{P(V=0|Y=0)P(V^*=1|V=0,Y=1)\}^{n_{03}} \times \\
& \{P(V=0|Y=0)P(V^*=0|V=0,Y=0)\}^{n_{04}} \times \\
& \{P(V=1|Y=1)P(V^*=0|V=1,Y=1)\}^{n_{11}} \times \\
& \{P(V=1|Y=1)P(V^*=1|V=1,Y=1)\}^{n_{12}} \times \\
& \{P(V=0|Y=1)P(V^*=1|V=0,Y=1)\}^{n_{13}} \times \\
& \{P(V=0|Y=1)P(V^*=0|V=0,Y=1)\}^{n_{14}} \times \\
& \{P(V^*=1|Y=0)\}^{n_{05}} \{1-(P(V^*=1|Y=0))\}^{n_{06}} \times \\
& \{P(V^*=1|Y=1)\}^{n_{15}} \{1-(P(V^*=1|Y=1))\}^{n_{16}} \\
=\ & \{r_0 SN\}^{n_{01}} \{r_0(1-SN)\}^{n_{02}} \{(1-r_0)(1-SP)\}^{n_{03}} \times \\
& \{(1-r_0)SP\}^{n_{04}} \{r_1 SN\}^{n_{11}} \{r_1(1-SN)\}^{n_{12}} \times \\
& \{(1-r_1)(1-SP)\}^{n_{13}} \{(1-r_1)SP\}^{n_{14}} \times \\
& \{r_0 SN + (1-r_0)(1-SP)\}^{n_{05}} \times \\
& \{1-(r_0 SN + (1-r_0)(1-SP))\}^{n_{06}} \times \\
& \{r_1 SN + (1-r_1)(1-SP)\}^{n_{15}} \times \\
& \{1-(r_1 SN + (1-r_1)(1-SP))\}^{n_{16}}. \qquad (2.2)
\end{aligned}
$$

This likelihood does not lead to a closed form for the maximum likelihood estimates of r_0, r_1, SN and SP. In quasi-Newton methods, the Hessian matrix of second derivatives of the function to be optimized is not required. That is why, a general-purpose optimization based on quasi-Newton methods or a variable metric algorithm is used to optimize Equation (2.2), specifically the algorithm that was published simultaneously in 1970 by *Broyden* [1970], *Fletcher* [1970], *Goldfarb* [1970] and *Shanno* [1970] (that is the origin of the name Broyden - Fletcher - Goldfarb - Shanno or BFGS method). This algorithm uses function values and gradients to build up a picture of the surface to be optimized.

However, for differential misclassification, we do have closed form expression for the maximum likelihood estimates of r_0, r_1, SN and SP.

2.3 Variance Estimates

The numerical approximation to the Hessian matrix can be obtained from the BFGS algorithm (implemented in `optim` function of `R`). The negative of the Hessian matrix is the observed Fisher information matrix. The inverse of the observed Fisher information matrix yields the asymptotic covariance matrix of the maximum likelihood estimates. By the use of multivariate delta method, one can easily obtain the asymptotic variance of the log odds ratio, given the estimated prevalence rates.

2.4 Likelihood-ratio Tests

A likelihood-ratio test is a statistical test for making a decision between two hypotheses based on the value of the ratio of the likelihood under two different hypotheses. The null hypothesis is often stated by saying the parameter ϕ is in a specified subset Φ_0 of the parameter space Φ. The likelihood function is $L(\phi) = L(\phi|\mathbf{x})$ is a function of the parameter ϕ with $\mathbf{x}$ held fixed at the value that was actually observed, i.e., the data. The likelihood ratio is

$$\Lambda = \frac{\sup L(\phi|\mathbf{x}) : \phi \in \Phi_o}{\sup L(\phi|\mathbf{x}) : \phi \in \Phi}.$$

The numerator corresponds to the maximum likelihood of the observed result under the null hypothesis H_0. The denominator corresponds to the maximum likelihood of the observed result under the alternative hypothesis H_1. Lower values of the likelihood ratio mean that the observed result was less likely to occur under the null hypothesis. Higher values mean that the observed result was more likely to occur under the null hypothesis. The likelihood ratio Λ is between 0 and 1. The likelihood ratio test rejects the null hypothesis if Λ is less than a critical value which is chosen to obtain a specified significance level α. Usually it is difficult to determine the exact distribution of the likelihood ratio for a specific problem. However, as the sample size n approaches ∞, the test statistic $-2\log(\Lambda)$ will be asymptotically χ^2 distributed with degrees of freedom equal to the difference in dimensionality of Φ_0 and Φ. In the current context, for without validation data, $\Phi_0 = \theta$ and $\Phi = (\theta_0, \theta_1)$. Similarly, for with validation data, $\Phi_0 = (r, SN, SP)$ and $\Phi = (r_0, r_1, SN, SP)$. Eventually from these tests, we obtain p-values.

A convenient measure of the performance of any hypothesis test is to find the probability of not making type II errors $(1 - \beta)$, or in other words, not

making the error of "not rejecting null hypothesis when it is false" - power of the test. Powers can be thought as the ability of the hypothesis test to detect a false null hypothesis. In Chapter 4, we will use the power curve as a tool to compare the tests based on with and without validation data. Also we will try to identify whether frequentist methods perform better than Bayesian methods or not. We will discuss relevant Bayesian methodology in Chapter 3.

Chapter 3

Bayesian Adjustment

3.1 Introduction

3.1.1 Bayes' Theorem

Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities of random events. For the random variable $\mathbf{X}$, that is distributed as $L(\phi|\mathbf{x})$, where ϕ is the parameter of interest, let $f_{\mathbf{X}}(\mathbf{x})$ is the marginal distribution and hence a function of the observed $\mathbf{X}$ alone, while $g(\phi)$ is the distribution of ϕ before observing $\mathbf{X}$. Then Bayes' Theorem says that the form of posterior distribution is:

$$\begin{aligned} \pi(\phi|\mathbf{x}) &= \frac{f(\mathbf{x},\phi)}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{g(\phi)L(\phi|\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \\ &\propto g(\phi)L(\phi|\mathbf{x}). \end{aligned} \tag{3.1}$$

Although Equation (3.1) seems simple, it is a fundamental theorem which has deep impact in statistical theory. It is often the case that the posterior $\pi(\phi|\mathbf{x})$ is non-standard or high dimensional, involving a lot of parameters. Then it is difficult to evaluate summaries such as the mean, variance, moments, etc. which require integration. Although analytically this is a difficult problem, algorithms discussed in the following sections help us find solutions numerically. It should be noted that simpler methods such as Laplace approximation can also be used to evaluate such summary quantities, but they require restrictive assumptions such as normal approximation to the posterior distribution and so on. Therefore, we consider algorithms that can be applied in broader contexts.

3.2 MCMC Algorithms

From §2.4, frequentist likelihood ratio test results are based on the asymptotic assumption, that is, a χ^2 approximation for the sampling distribution

of the test statistic $-2\log(\Lambda)$, since the exact distribution of the statistic is hard to determine and varies from problem to problem. Monte Carlo methods can be an alternative to this approach. These methods can even be applied in the cases where the distributions are not in conventional format or unknown.

To explain the idea of Monte Carlo, suppose that ϕ is a collection of model parameters or unknowns, and $h(\phi)$ is a function of ϕ. We want to evaluate the expected value of the given function $h(\phi)$ over a pdf $\pi(\phi)$. In other words, we want to evaluate $E_\pi\big(h(\phi)\big) = \int h(\phi)\pi(\phi)d\phi$. If π has a very complex form, we proceed with the Monte Carlo integration technique. Here, we draw samples $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(n)}$ independently from $\pi(\phi)$. Then we estimate $\widehat{E}_\pi\big(h(\phi)\big) = \frac{1}{n}\sum_{i=1}^{n} h(\phi^{(i)})$, which can be made as accurate as desired by increasing sample size. Therefore, the fundamental idea behind Monte Carlo methods is that, by repeatedly drawing random samples from the target population $\pi(\phi)$, we can gain insight regarding the behavior of a statistic. When we observe the behavior for a very long time, we obtain an estimate of the sampling distribution of the statistic. But this added advantage is not without a price - time and computer resources are big issues for these algorithms. However, recent advances in computing technologies have led to enormous popularity of Monte Carlo simulation as a powerful alternative to formula-based analytic approaches, especially where the solution requires a lot of assumptions.

In the Bayesian context, this $\pi(\phi)$ is the posterior density $\pi(\phi|\mathbf{x})$, which may have a nonstandard, complicated form. Here $\mathbf{x}$ denotes the observed information, and ϕ is high dimensional. Sampling independently from the posterior density $\pi(\phi|\mathbf{x})$ is generally not feasible, and closed form solutions are not usually possible. Therefore, we generate a chain or dependent samples $\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(n)}$ from the posterior using a Markov Chain scheme. This Markov chain generates each iteration $\phi^{(i)}$ taking into account of the previous value $\phi^{(i-1)}$ only. We want to create a Markov Chain whose stationary or limiting or equilibrium distribution is the desired posterior $\pi(\phi|\mathbf{x})$. Here the posterior distribution $\pi(\phi|\mathbf{x})$ is the target distribution. To obtain the stationary distribution of the Markov Chain, we need to run the burn-in for a long time. Here, burn-in refers to the series of initial samples that are not expected to have yet converged to the target distribution and are hence excluded from any subsequent analysis. In brief, the basic idea of Markov Chain Monte Carlo is to iteratively produce parameter values that are representative samples from the joint posterior. For large number of iterations,

this scheme will provide samples $\phi^{(i)}$ from its stationary distribution, that is, when the successive samples becomes uncorrelated. This way, it is surprisingly easy to approximate the posterior distribution $\pi(\phi|\mathbf{x})$. However, one added disadvantage to this entire procedure is that we have to monitor convergence. We will discuss this further in §3.3.

The Markov Chain Monte Carlo algorithms that are used in the Bayesian version of the test under consideration are described in §3.2.3. But for detailed understanding of the procedure, we start with a general description of the Metropolis-Hastings algorithm and Gibbs algorithm. However, for basic terminologies and definitions used in these Markov Chain Monte Carlo algorithms, we refer the readers to *Gelman* [2004].

3.2.1 Metropolis-Hastings Algorithm

Suppose we need to estimate a parameter vector with k-elements, $\phi \in \Phi$ and the posterior, $\pi(\phi)$. When the chain reaches the position ϕ at the t^{th} step, we draw ϕ' from a distribution over the same support and we name it the proposal or jumping distribution, $P_t(\phi'|\phi)$, according to which a new value ϕ' (candidate point) is proposed given the new current value $\phi^{[t]}$. One thing to keep in mind is that $P_t(\phi'|\phi)$ should be easy to sample from. We are producing a multidimensional candidate value. The condition here is that the reverse function value, $P_t(\phi|\phi')$ should also exist. In the literature, the acceptance ratio is defined as follows:

$$\alpha(\phi', \phi) = \frac{\pi(\phi')P_t(\phi|\phi')}{\pi(\phi)P_t(\phi'|\phi)} \tag{3.2}$$

The Metropolis-Hastings algorithm does not necessarily move on every iteration. The probabilistic rule that decides whether the candidate point is accepted or not, i.e., transition from t to $(t+1)$th point, is:

$$\phi^{[t+1]} = \begin{cases} \phi' & \text{with probability} \quad \min\{\alpha(\phi', \phi^{[t]}), 1\} \\ \phi^{[t]} & \text{with probability} \quad 1 - \min\{\alpha(\phi', \phi^{[t]}), 1\} \end{cases}$$

We only need to know the posterior distribution $\pi(\phi)$ up to a constant of proportionality. This is considered as the most attractive feature of the Metropolis-Hastings sampler.

A single Metropolis-Hastings iteration proceeds with the following steps:

1. Initialize the chain with any arbitrary value.

2. Generate a candidate point ϕ' from $P(\phi'|\phi)$, where ϕ is the current location.

3. Sample u from uniform $(0, 1)$ distribution.

4. If $\alpha(\phi'|\phi) \geq u$ then accept ϕ'.

5. Otherwise keep ϕ as the new location and repeat until convergence.

From the obtained chain, we truncate burn-in samples, and the rest of the chain is used to estimate the posterior distribution.

3.2.2 Gibbs Algorithm

The Gibbs sampler is a special case of Metropolis-Hastings where we always accept a candidate value. The idea of Gibbs sampling is that, given a multivariate distribution, sample from a conditional distribution. This sampling is generally simpler than integrating over a joint distribution. Hence, the Gibbs sampler is simply a Markovian updating scheme, based on a sequence of conditional probabilistic statements.

We will give a brief outline of Gibbs algorithm in its simplest form. Let the joint distribution of interest be $\pi(\phi)$, where ϕ is a vector of of k parameters. The aim is to create a Markov chain that cycles through some conditional statements. A requirement for use of this sampler is that we must know the full conditional distributions. This is a major limitation of this algorithm, especially for the cases where the conditional distributions are hard to derive. The full set of required conditional distributions for ϕ are denoted by Φ and defined by $\pi(\Phi) = \pi(\phi_i|\phi_1, \phi_2, \ldots \phi_{i-1}, \phi_{i+1}, \ldots \phi_k)$ for $i = 1, \ldots, k$. It should be possible to draw samples from these conditional distributions. At each iteration of the Gibbs sampling, the algorithm cycles through these conditionals based on the most recent version of all other parameters. The order is not important, but it is important that the most recent draws from the other samples be used. The algorithm is as follows:

1. Decide on the starting values: $\phi^{[0]} = [\phi_1^{[0]}, \phi_2^{[0]}, \ldots, \phi_k^{[0]}]$.

2. At the t^{th} iteration, a single cycle is completed by drawing values from

the following k distributions:

$$
\begin{aligned}
\phi_1^{[t]} &\sim \pi(\phi_1 \mid \phi_2^{[t-1]}, \phi_3^{[t-1]}, \phi_4^{[t-1]}, \ldots, \phi_{k-1}^{[t-1]}, \phi_k^{[t-1]}) \\
\phi_2^{[t} &\sim \pi(\phi_2 \mid \phi_1^{[t]}, \phi_3^{[t-1}, \phi_4^{[t-1}, \ldots, \phi_{k-1}^{[t-1]}, \phi_k^{[t-1}) \\
\phi_3^{[t} &\sim \pi(\phi_3 \mid \phi_1^{[t]}, \phi_2^{[t]}, \phi_4^{[t-1]}, \ldots, \phi_{k-1}^{[t-1]}, \phi_k^{[t-1]}) \\
&\vdots \\
\phi_{k-1}^{[t]} &\sim \pi(\phi_{k-1} \mid \phi_1^{[t]}, \phi_2^{[t}, \phi_3^{[t]}, \ldots, \phi_{k-2}^{[t]}, \phi_k^{[t-1]}) \\
\phi_k^{[t]} &\sim \pi(\phi_k \mid \phi_1^{[t]}, \phi_2^{[t]}, \phi_3^{[t]}, \ldots, \phi_{k-2}^{[t]}, \phi_{k-1}^{[t}).
\end{aligned}
$$

Here ϕ_i can be a multidimensional vector.

3. Set $t = (t+1)$ and repeat until convergence.

If the Gibbs sampler has run for sufficiently long time, it produces samples from the desired stationary distribution. The attractive feature of the Gibbs sampling algorithm is that these conditional distributions contain enough information to eventually produce samples from the desired joint distribution.

3.2.3 Mixed Algorithm

With Validation Data

Likelihood Function: First, we define the parameter space:

$$\Omega \equiv (r_0, r_1, SN, SP),$$

where r_0 is the exposure prevalence for controls and r_1 is the same for cases, SN is the sensitivity and SP is the specificity under nondifferential classification.

The cell counts u_{ij} of the main data as shown by Table 1.1 are generated from a binomial distribution. To be more specific, the actual number of subjects that are in positive exposure status (u_{i1}) amongst those who are exposed in the groups of cases or controls (n_{i5}) follows a binomial with parameters n_{i5} and PPV_i (as defined in Equation (1.4)). Likewise, conditioning on the number of cases or controls with negative exposure status (n_{i6}), the number of truly unexposed subjects (u_{i4}) follows a binomial with parameters n_{i6} and NPV_i (as defined in Equation (1.5)).

The likelihood function for this setting is given in Equation (3.3). Here, the data $\tilde{Y}$ is updated as $\tilde{Y} = \{Y_n, Y_u\} = \{(n_{i1}, n_{i2}, n_{i3}, n_{i4}), (u_{i1}, u_{i2}, u_{i3},$

$u_{i4})\ \}_{i=0}^{1}$.

$$
\begin{aligned}
f(Y_n, Y_u|\Omega) &= L(r_0, r_1, SN, SP|Y_n, Y_u) \\
&\propto \prod_{i=0}^{1}\Big[\{P(V^*=1|V=1,Y=i)P(V=1|Y=i)\}^{n_{i1}} \\
&\quad \times \{P(V^*=0|V=1,Y=i)P(V=1|Y=i)\}^{n_{i2}} \\
&\quad \times \{P(V^*=1|V=0,Y=i)P(V=0|Y=i)\}^{n_{i3}} \\
&\quad \times \{P(V^*=0|V=0,Y=i)P(V=0|Y=i)\}^{n_{i4}} \\
&\quad \times \{P(V^*=1|V=1,Y=i)P(V=1|Y=i)\}^{u_{i1}} \\
&\quad \times \{P(V^*=0|V=1,Y=i)P(V=1|Y=i)\}^{u_{i2}} \\
&\quad \times \{P(V^*=1|V=0,Y=i)P(V=0|Y=i)\}^{u_{i3}} \\
&\quad \times \{P(V^*=0|V=0,Y=i)P(V=0|Y=i)\}^{u_{i4}}\Big] \\
&= \prod_{i=0}^{1}\Big[\{SN_i r_i\}^{n_{i1}} \times \{(1-SN_i)r_i\}^{n_{i2}} \times \{(1-SP_i)(1-r_i)\}^{n_{i3}} \\
&\quad \times \{SP_i(1-r_i)\}^{n_{i4}} \times \{SN_i r_i\}^{u_{i1}} \times \{(1-SN_i)r_i\}^{u_{i2}} \\
&\quad \times \{(1-SP_i)(1-r_i)\}^{u_{i3}} \times \{SP_i(1-r_i)\}^{u_{i4}}\Big] \\
&= \prod_{i=0}^{1}\Big[\{SN_i r_i\}^{n_{i1}+u_{i1}} \times \{(1-SN_i)r_i\}^{n_{i2}+u_{i2}} \\
&\quad \times \{(1-SP_i)(1-r_i)\}^{n_{i3}+u_{i3}} \times \{SP_i(1-r_i)\}^{n_{i4}+u_{i4}}\Big] \\
&= \prod_{i=0}^{1}\Big[r_i^{n_{i1}+n_{i2}+u_{i1}+u_{i2}} \times (1-r_i)^{n_{i3}+n_{i4}+u_{i3}+u_{i4}} \times SN_i^{n_{i1}+u_{i1}} \\
&\quad \times (1-SN_i)^{n_{i2}+u_{i2}} \times (1-SP_i)^{n_{i3}+u_{i3}} \times SP_i^{n_{i4}+u_{i4}}\Big]. \qquad (3.3)
\end{aligned}
$$

Under nondifferential misclassification, $SN_0 = SN_1 = SN$ and $SP_0 = SP_1 = SP$ (according to the definition that we used). Therefore, in Equation (3.3), we could have used SN and SP, instead of SN_i and SP_i. But we preferred to keep the general format to present the likelihood function for the broader context.

Prior Specification: We are interested in $\Omega \equiv (r_0, r_1, SN, SP)$, as defined in section 3.2.3. Each of these parameters can possibly range from 0 to 1. To cover the whole real line from $-\infty$ to ∞, we make a logit transformation of each of these parameters. To keep the problem manageable, we

assume the following:

$$\begin{aligned}
\begin{pmatrix} \Pi_0 \\ \Pi_1 \end{pmatrix} &\equiv \begin{pmatrix} log\frac{r_0}{1-r_0} \\ log\frac{r_1}{1-r_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}\right), \\
\Gamma &\equiv \left(log\frac{SN}{1-SN}\right) \sim N\left(\mu_2, \sigma_2^2\right), \\
\Upsilon &\equiv \left(log\frac{SP}{1-SP}\right) \sim N\left(\mu_3, \sigma_3^2\right), \qquad (3.4)
\end{aligned}$$

where Π_0, Π_1, Γ, Υ are just the logit transformed versions of r_0, r_1, SN, SP respectively. Here $(\Pi_0, \Pi_1)'$ is assumed to follow a bivariate normal distribution with hyperparameters μ_0, μ_1 and σ_0, σ_1, ρ. Similarly, Γ and Υ follow independent normals with hyperparameters μ_2, σ_2 and μ_3, σ_3.

Also conditional distribution of a bivariate normal variable remains normal, therefore, given Γ, Υ, we have

$$\begin{aligned}
\Pi_0|\Pi_1 &\sim N\left(\{\mu_0 + \rho\frac{\sigma_0}{\sigma_1}(\Pi_1 - \mu_1)\}, \sigma_0^2(1-\rho^2)\right) \\
\Pi_1|\Pi_0 &\sim N\left(\{\mu_1 + \rho\frac{\sigma_1}{\sigma_0}(\Pi_0 - \mu_0)\}, \sigma_1^2(1-\rho^2)\right)
\end{aligned}$$

It should be noted that these assumptions of independence among the parameters and normal distribution structures of them are purely based on mathematical convenience. Researchers can think of other possible distributions if they find them suitable for the purpose. Also if one thinks that the assumption of independence of the parameters is inappropriate, it is possible to impose correlation among the parameters by means of some multivariate distribution with defined correlation structure.

We assume that the analyst's prior beliefs about the logit transformed parameters can be represented by the hyperparameters mentioned in Equation (3.4). These beliefs may be gained from relevant examples from the given subject area. Under fairly general conditions, we have empirical reason to believe that both r_0 and r_1 usually lie between $r_{min} = 0.02$ and $r_{max} = 0.50$; we will assume a median being 0.125. Then $\mu = \mu_0 = \mu_1 = 0.125$. Within 2 standard deviation, on logit scale, we have $\sigma_0 = \sigma_1 = \{logit(\mu) - logit(r_{min})\}/3$ under normality with 95% probability. Also assume a mild value for ρ, say, 0.3 to allow relatively large standard deviation of $\log OR$ around the mean of 0. For SN and SP, we usually see them lying

between 0.60 and 0.99; we will assume median of 0.80. Using the same logic as before, we determine the hyperparameters. It should be emphasized that the user can choose any hyperparameters of interest. The above is just an example of how we can construct the prior from the mentioned empirical beliefs. Often the posterior is robust to the assumed prior. We will discuss this point further in Chapter 5.

Posterior: Since Equation (3.3) is a complex one, simulating Ω directly from the joint posterior distribution is troublesome. Therefore, we will sample sequentially from the conditional distributions as follows:

$$\begin{aligned}
\pi(r_0, r_1|\tilde{Y}, SN, SP) &\propto f_r(r_0, r_1) \prod_{i=0}^{1} \left[r_i^{n_{i1}+n_{i2}+u_{i1}+u_{i2}} (1-r_i)^{n_{i3}+n_{i4}+u_{i3}+u_{i4}} \right], \\
\pi(SN|\tilde{Y}, r_0, r_1, SP) &\propto f_{SN}(SN) \prod_{i=0}^{1} \left[SN_i^{n_{i1}+u_{i1}} \times (1-SN_i)^{n_{i2}+u_{i2}} \right], \\
\pi(SP|\tilde{Y}, SN, r_0, r_1) &\propto f_{SP}(SP) \prod_{i=0}^{1} \left[SP_i^{n_{i4}+u_{i4}} \times (1-SP_i)^{n_{i3}+u_{i3}} \right], \qquad (3.5)
\end{aligned}$$

using the prior distribution f_r, f_{SN} and f_{SP} as already described.

Since the densities are not conditionally conjugate, we implement univariate Metropolis-Hastings jumps embedded in the Gibbs sampling. This algorithm will update each component in the pairs of parameters, (r_0, r_1), and the same for SN and SP. For satisfactory performance of the MCMC, we need to make suitable choice of jumping distribution. If we examine the likelihood function in Equation (3.3), and think of r_i, SN and SP separately, it looks similar to a beta density. Hence we assume a beta jumping distribution. This simplifies calculation of the acceptance rate by cross canceling the ratio of proposed versus current likelihoods and the ratio between two jumping densities. For example, consider the acceptance probability for the one-dimensional M-H jump on r_0 in Equation (3.5). The jumping rule is specified as $r_0' \sim Beta(n_{01}+n_{02}+u_{01}^t+u_{02}^t+1, n_{03}+n_{04}+u_{03}^t+u_{04}^t+1)$, close to the conditional sampling distribution, where t is the index for iteration

number. The ratio in Equation (3.2) becomes

$$\begin{aligned}
\alpha &= \frac{\frac{\pi(r_0'|r_1^t,SN^t,SP^t,Y^t)}{\pi(r_0^t|r_1^t,SN^t,SP^t,Y^t)}}{\frac{P_t(r_0'|r_0^t,r_1^t,SN^t,SP^t)}{P_t(r_0^t|r_0',r_1^t,SN^t,SP^t)}} \\
&= \frac{\frac{\pi(r_0'|r_1^t,SN^t,SP^t)L(r_0',r_1^t,SN^t,SP^t)}{\pi(r_0^t|r_1^t,SN^t,SP^t)L(r_0^t,r_1^t,SN^t,SP^t)}}{\frac{L(r_0',r_1^t,SN^t,SP^t)}{L(r_0^t,r_1^t,SN^t,SP^t)}} \\
&= \frac{\pi(r_0'|r_1^t,SN^t,SP^t)}{\pi(r_0^t|r_1^t,SN^t,SP^t)} \\
&= \frac{\pi(r_0'|r_1^t)}{\pi(r_0^t|r_1^t)}.
\end{aligned}$$

Therefore, we are left with merely the ratio between two prior distributions. Thus, using this mixed algorithm, we proceed as follows:

1. Set starting values of $(r_0^0, r_1^0, SN^0, SP^0)$.

2. At the t^{th} iteration,

   - Given parameters $\Omega^t = (r_0^t, r_1^t, SN^t, SP^t)$, generate new data as unobserved actual exposure data $Y_u^t = \{u_{ij}^t\}$ based on binomial distributions, for $i = 0, 1$, $j = 1, 2, 3, 4$.
   - Based on the updated cell counts $\{u_{ij}^t\}$ at the t^{th} iteration, model parameters are generated as follows:

     (i) Simulate r_0^t conditioning on $(r_1^{t-1}, SN^{t-1}, SP^{t-1})$ using the M-H algorithm. the proposed jumping rule is $r_0' \sim Beta(n_{01}+n_{02}+u_{01}^t, n_{03}+n_{04}+u_{03}^t)$, with acceptance rate $\min\left\{\frac{\pi(r_0'|r_1^{t-1})}{\pi(r_0^{t-1}|r_1^{t-1})}\right\}$.

     (ii) Simulate r_1^t conditioning on $(r_0^t, SN^{t-1}, SP^{t-1})$ using the M-H algorithm. The proposed jumping rule is $r_1' \sim Beta(n_{11}+n_{12}+u_{11}^t+u_{12}^t+1, n_{13}+n_{14}+u_{13}^t+u_{14}^t+1)$, with acceptance rate $\min\left\{\frac{\pi(r_1'|r_0^t)}{\pi(r_1^{t-1}|r_0^t,1)}\right\}$.

     (iii) Simulate SN^t conditioning on (r_0^t, r_1^t, SP^{t-1}) using the M-H algorithm. The proposed jumping rule is $SN' \sim Beta(n_{01}+u_{01}^t+n_{11}+u_{11}^t+1, n_{02}+u_{02}^t+n_{12}+u_{12}^t+1)$, with acceptance rate $\min\left\{\frac{\pi(SN'|SN^{t-1})}{\pi(SN^{t-1}|SN^{t-1})}\right\}$.

     (iv) Simulate SP^t conditioning on (r_0^t, r_1^t, SN^t) using the M-H algorithm. The proposed jumping rule is $SP' \sim Beta(n_{04}+$

$u_{04}^t + n_{14} + u_{14}^t + 1, n_{03} + u_{03}^t + n_{13} + u_{13}^t + 1)$, with acceptance rate $\min\left\{\frac{\pi(SN|SP^{t-1})}{\pi(SP^{t-1}|SP^{t-1})}\right\}$.

- Calculate the log odds ratio ϕ at the t^{th} iteration.

3. Repeat the step (2) at subsequent iterations, for $t = 1, \ldots, m+n$, to simulate target parameters using the hybrid algorithm.

The procedure is run for sufficiently long $m+n$ iterations, where m is the number of burn-in iterations and n is the number of target iterations.

Without Validation Data

Likelihood Function: At first, we define the parameter space:

$$\tilde{\Omega} \equiv (\theta_0, \theta_1),$$

where θ_0 is the apparent exposure prevalence for controls and θ_1 is the same for cases.

As for validation data case, let us assume that the cell counts u_{ij} from the main data as shown by Table 1.1 are generated from a binomial distribution.

The likelihood function from this setting becomes

$$\begin{aligned} L(\tilde{\Omega} = \{\theta_0, \theta_1\}|Y_n, Y_u) &\propto \prod_{i=0}^{1} \theta_i^{n_{i1}+n_{i3}+n_{i5}}(1-\theta_i)^{n_{i2}+n_{i4}+n_{i6}} \\ &= \theta_0^{n_{01}+n_{03}+n_{05}}(1-\theta_0)^{n_{02}+n_{04}+n_{06}} \times \\ &\quad \theta_1^{n_{11}+n_{13}+n_{15}}(1-\theta_0)^{n_{12}+n_{14}+n_{16}} \end{aligned}$$

Prior Specification: We are interested in $\Omega \equiv (\theta_0, \theta_1)$. Each of these parameters ranges from 0 to 1. To cover the whole real like from $-\infty$ to ∞, we make a logit transformation of each of these parameters. To keep the problem manageable, we assume the following:

$$\begin{pmatrix} \Theta_0 \\ \Theta_1 \end{pmatrix} \equiv \begin{pmatrix} log\frac{\theta_0}{1-\theta_0} \\ log\frac{\theta_1}{1-\theta_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \tilde{\mu}_0 \\ \tilde{\mu}_1 \end{pmatrix}, \begin{pmatrix} \tilde{\sigma}_0^2 & \tilde{\rho}\tilde{\sigma}_0\tilde{\sigma}_1 \\ \tilde{\rho}\tilde{\sigma}_0\tilde{\sigma}_1 & \tilde{\sigma}_1^2 \end{pmatrix}\right),$$

where Θ_0, Θ_1 are just the logit transformed versions of θ_0, θ_1 respectively. Here (Θ_0, Θ_1) is assumed to follow bivariate normal distribution with hyperparameters $\tilde{\mu}_0, \tilde{\mu}_1$ and $\tilde{\sigma}_0, \tilde{\sigma}_1, \tilde{\rho}$. For the hyperparameters, the logic is the same as for the prior of (Π_0, Π_1) in §3.2.3. The conditional distributions of a bivariate normal can be similarly derived.

Posterior: Similar to the case of data with a validation sub-sample, we proceed as follows for the mixed algorithm:

1. Set starting values of (θ_0^0, θ_1^0).
2. At the t^{th} iteration,
   - Given parameters $\tilde{\Omega}^t = (\theta_0^t, \theta_1^t)$, update the unobserved actual exposure data.
   - Based on the updated cell counts $\{u_{ij}^t\}$ at the t^{th} iteration, model parameters are generated as follows:
     (i) Simulate θ_0^t conditioning on θ_1^{t-1} using the M-H algorithm. The proposed jumping rule is $\acute{\theta_0} \sim Beta(n_{01} + n_{03} + n_{05} + 1, n_{02}+n_{04}+n_{06}+1)$, with acceptance rate $\min\left\{\frac{\pi(\acute{\theta_0}|\theta_1^{t-1})}{\pi(\theta_0^{t-1}|\theta_1^{t-1})}\right\}$.
     (ii) Simulate θ_1^t conditioning on θ_0^t. The proposed jumping rule is $\acute{\theta_1} \sim Beta(n_{11} + n_{13} + n_{15} + 1, n_{12} + n_{14} + n_{16} + 1)$, with acceptance rate $\min\left\{\frac{\pi(\acute{\theta_1}|\theta_0^t)}{\pi(\theta_1^{t-1}|\theta_0^t,1)}\right\}$.
   - Calculate the log odds ratio ϕ at the t^{th} iteration.
3. Repeat step (2) at subsequent iterations, for $t = 1, \ldots, m + n$, to simulate target parameters alternately using the hybrid algorithm.

3.3 MCMC Diagnostics

Formal convergence diagnostic techniques are addressed here, to identify various frequently occurring issues regarding mixing and coverage of the MCMC algorithms discussed in §3.2. There are several common issues as discussed by *Gill* [2008]:

- There is no formal way to ensure that the chain at currently in the target distribution for a given Markov chain at a given time.
- It is not possible to ensure that a Markov chain will explore all areas of the target distribution in finite time.
- Slow convergence. Although theoretically this is not a problem, it is a practical issue.

Fundamentally these concerns can be summarized as setting up the parameters of the process appropriately, ensuring satisfactory mixing throughout

the whole sample space, and obtaining convergence at some point. There are some design issues that must be taken into consideration before constructing and running the chain. Some of these considerations are taking decisions like determining where to start the chain, judging how long to burn-in the chain before recording values for inference, and determining whether to thin the chain values by discarding portions of the output.

1. *Initialization:* When little is known about the process, some researchers randomly distribute initial values through the state space. Usually it is best to try several different starting points in the state space and observe whether they lead to noticeably different descriptions of the posteriors. This is an obvious sign of non-convergence of the Markov chain. Unfortunately the reverse is not true: it is not the case that if one starts several Markov chains in different places in the state space and they congregate for a time in the same region that this is the region that characterizes the stationary distribution. It is possible that all of the chains are influenced by the same local maxima.

2. *Burn-In:* The beginning set of runs are discarded as they are not expected to be representative of the target distribution. Unfortunately, there is no formal way to calculate the appropriate length of the burn-in period. Assessing diagnostic plots or other convergence statistics described in the literature are the usual ways to determine the burn-in period.

3. *Mixing:* A chain that has not fully explored the stationary distribution will tend to give biased results since it is based on only a subset of the state space. Often slow mixing through the target distribution can be attributed to high correlation between model parameters - hence checking autocorrelation is a good idea. This is particularly the case with the Gibbs sampling algorithm. High intra-parameter correlation is also an issue with the Metropolis-Hastings algorithm since it will also induce slow mixing, due to observing too many rejected candidate values.

4. *Chain thinning:* In the very long simulations, storage of the observed values on the computer becomes a huge problem. Not only the storage, but also the process of storing the high dimensional parameter realizations will slow down the computation. The idea of thinning the chain is to run the chain in an usual fashion, but record only every c-th value of the chain, thus reducing the storage demands while still preserving the general trend of the Markov process. Here c is some small

integer. It is worth mentioning that this approach does not improve the quality of the estimate, speed up the chain or help in convergence - rather it is the other way around - the variance estimate will be somewhat distorted due to use of less observations. Still, it is a tool for dealing with possibly limited computer resources. Given the tradeoffs between storage and accuracy as well diagnostic ability, the value of c should be carefully chosen in any given problem.

Keeping all the above aspects in mind, we still need to find the number of iterations that would be sufficient for approximating the convergence to the target distribution or the length of burn-in sample. Various methods are proposed in the literature for monitoring the convergence of Markov Chain Monte Carlo chains; see *Cowles and Carlin* [1996], *Brooks et al.* [1997], *Geyer* [1992], *Raftery and Lewis* [1992], *Hastings* [1970], *Robert* [1995] and *Rizzo* [2007] for more detailed discussion. We will discuss the graphical diagnosis and the approach suggested by *Gelman and Rubin* [1992] and gel; *Gelman* [2004].

3.3.1 Conventional Graphical Diagnosis

Graphically, trace plots are the most popular way to assess convergence. If the iterations are run for fairly long time, the trend will move from initial values to the desired density. Other plots that are popularly used include the mean graph - which plots the mean scores of the previous values versus the iteration number. If the chain is stable, a flat line will be produced. This does not evidently prove convergence, but if the chain is not producing a flat line, this indicates that the chain has not yet converged. Also, density plots of the estimates after burn-in can be drawn.

3.3.2 Gelman-Rubin Method for Monitoring Convergence

gel suggested that the lack of convergence can be appropriately detected by comparing multiple sequences (at least two) with initial points being widely dispersed in the target distribution. The Gelman-Rubin statistic $\hat{R}$ (shrink factor) of monitoring convergence of a Markov chain is based on comparing the behavior of a group of chains with respect to the variance of a given scalar summary statistic. The estimates of the variance of the statistic are analogous to estimates based on between-sample and within-sample mean squared errors in a one-way analysis of variance. It uses the between sequence variation of the summary statistic as an upper bound and the within-sequence variance as a lower bound. The idea behind this

method is that, if the chain converges to the target distribution, both the variances will also converge. It is recommended that the sequence be run until $\hat{R}$ for all the summaries are less than 1.2 at most. If it is less than 1.1, the convergence is even better.

Chapter 4

Simulation Results

4.1 Data Generation

To generate data for our setting as described in Table 1.1 and Table 1.2, the steps are:

1. We independently generate the true exposure status ($V = 0$ and 1) at $Y = i$ for $i = 0, 1$. The generating distribution is Bernoulli with parameter r_i.

2. We generate surrogate measurements $V^*|V$ from Bernoulli based on the fact that

$$\begin{aligned} P(V^* = 1|V = 0) &= 1 - SN \\ P(V^* = 1|V = 1) &= SP, \end{aligned}$$

where r_i is the exposure prevalence, SN is the sensitivity and SP is the specificity under nondifferential classification. Now we cross-tabulate to get the validation table. The main data generation is exactly the same - but in this case we omit the true exposure status (V) from the classification - it is only about apparent measurements.

4.2 Scenario Settings Under Frequentist Adjustment

While dealing with frequentist adjustments, we utilize all the n's (observed values) but not the u's (unobserved values) as mentioned in Tables 1.1 and 1.2. In the model without validation data or the two parameter model (these two names of this model will be used interchangeably throughout the entire work) discussed in §2.2.1, we simply use the column totals from the tables. But in the model with validation data (or four parameter model) discussed in §2.2.2, we also use the n's that are inside the validation table. Hence, when we make comparison, for example - say, for sample size 2000 where

200 are in the validation and 1800 are in the main (unvalidated) part, then for without validation data we use the 2000 subjects aggregately as if there were no validation part (marginal total for the surrogate variable are known for both parts and the without validation model uses just these marginal totals). But with the validation model, we can recognize 200 subjects as comprising the validation part and the rest as the main part.

To understand the performance of frequentist adjustment for nondifferential misclassification in the simplest possible way, several scenarios are considered, as shown in Table 4.1, varying the level of exposure prevalence, or sensitivity and specificity, or sample size, or sample proportion of the validation and main parts of the data. Notice that, the whole process is very complex. Here, the factors are merely assessed in an uni-dimensional way in all these cases, that is, all other factors are held constant when we switch from one scenario setting to another, so we will not be able to assess the possibility of interactions among the factors. That would require more combinations of scenarios and a huge amount of data would have to be generated. However, there would be some limitations to that approach as well - such as computing time and storage and, above all, comprehending and interpreting all those data would be challenging. As our objective of assessing impacts on hypothesis testing is a relatively new one in epidemiologic research, this simplified approach should provide some rough ideas about the effects which will suffice as a first step in the process.

We use power curves as the tool of comparison for this frequentist approach. Therefore, the null hypothesized value (difference between the exposure prevalences is zero) is the mid-value on the horizontal axis. On the right and left side of it, four other equidistant difference points are selected in each direction based on the difference of the exposure prevalences from case and control groups, according to alternative hypothesis. In this work, the considered absolute difference between exposure prevalences from case and control groups were 0.05, 0.10, 0.15, 0.20 (fixing r_0 and changing r_1 to achieve the desired difference). Therefore, we have nine points in total in one power curve. The process of getting the estimated power is as follows for any one point: 10,000 datasets are generated according to the hypothesized difference in exposure prevalence from two groups. We implement the hypothesis test on each dataset and evaluate the p-value. The estimated power is given by the proportion of the datasets that provide a p-value less than the chosen level of significance 0.05. In theory, with a large number of datasets, the lowest point of the power should be the chosen level of sig-

nificance at the null hypothesized point. This is under the assumption that asymptotic cut offs are accurate.

Power curves from the hypotheses of $H_0 : \theta_0 = \theta_1 = \theta$ (from without validation data) and $H_0 : r_0 = r_1 = r$ (from with validation data) are shown together in each graph because of their equivalence as described in §2.2. On a technical note, to allow reproducibility of the results, the seed is chosen arbitrarily and kept the same throughout the entire analysis.

Table 4.1: Scenarios under consideration

Factor changed	SN, SP				Sample Size			
Scenarios	A	B	C	D	E	F	G	B
Validated data	200	200	200	200	200	200	200	200
Unvalidated data	1800	1800	1800	1800	200	400	800	1800
$r_0 = r_1$	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
$SN = SP$	0.60	0.70	0.80	0.90	0.70	0.70	0.70	0.70
Factor changed	Exposure Prevalence				Proportion of data			
Scenarios	I	J	K	B	B	N	O	P
Validated data	200	200	200	200	200	500	1000	1500
Unvalidated data	1800	1800	1800	1800	1800	1500	1000	500
$r_0 = r_1$	0.25	0.30	0.35	0.40	0.40	0.40	0.40	0.40
$SN = SP$	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70

4.2.1 Under Different Values of Sensitivity and Specificity

We generated 10,000 datasets with exposure prevalence 0.4 for both case and control groups, where, in each dataset, 200 were in the validation part (50% in the case group, and the rest in the control group) of the data as was described in Table 1.2, and 1800 were in the main dataset (again 50% in the case group and the rest are in the control group) as was described in Table 1.1. Four different sets of sensitivity and specificity values were considered: 0.60, 0.70, 0.80 and 0.90. We implemented the likelihood ratio test (discussed in §2.2) for the two parameter (θ_0, θ_1) model for data without validation part and the four parameter (r_0, r_1, SN, SP) model for data with validation part. The estimated power curves out of these tests for all the cases under consideration are shown in Figure 4.1. From the figure, it is evident that the power of the two parameter model is always dominated by that of the four parameter model. The situation is much exacerbated

when the values of sensitivity and specificity are low. However, when the misclassification is at least 0.8, the powers of both tests are almost the same.

According to the theory, for the exact tests, the power at the null point (where the difference between the exposures of case and control groups are zero) should be equal to the level of significance, which is 0.05. Due to simulation error, this might deviate a bit. From the Figure 4.1, we can also see that the lowest point of powers do match at 0.05 in each setting. Therefore, the number of simulations considered here are adequate to show the power curves nicely.

4.2.2 Under Different Sample Sizes

Like the previous scenario, we generated 10,000 datasets. The exposure prevalence for both case and control groups was 0.4. Sensitivity and specificity of both groups was 0.7. The sample sizes varied in this scenario as follows: 400, 600, 1000 and 2000, where in each situations, we had 200 as the validation part of the data and rest were the main part of the data (again half are allocated in case group, and rest are in control group in each setting). Still the four parameter model is superior considering the power of the likelihood ratio tests, as shown in Figure 4.2. Naturally, as the sample size increases, the power of both the tests increases.

4.2.3 Under Different Exposure Prevalence Rates

In this scenario, we considered 10,000 datasets. Again sensitivity and specificity were set to be 0.7. In each dataset, we had 200 as the validation part of the data and 1800 as the main part of the data (50% in case group, and rest are in control group). The hypothesis regarding the exposure prevalence was always $H_0 : r_0 = r_1 = r$ or equivalently $H_0 : r_d = 0$, where $r_d = r_0 - r_1$. Alternative hypothesis in this case would be that the difference of r_0 and r_1 is not zero. To draw a complete power curve, we assume that the possible differences in horizontal axis are 0.05, 0.10, 0.15, 0.20 in both directions, so that we get nine points in total to draw a power curve. There were four values of r under consideration: $r = 0.25$, $r = 0.30$, $r = 0.35$ and $r = 0.40$. From Figure 4.3, in all the cases, the power of the two parameter model is less than the four parameter models, and the power does not seem to vary much under different exposure prevalence values r. In practical situations, sometimes we see much less prevalence. Hence, we construct power curves for lower prevalence rates such as $r = 0.005$, $r = 0.01$, $r = 0.05$ and $r = 0.10$,

Figure 4.1: Power curves under different sensitivity and specificity values: 0.6, 0.7, 0.8 and 0.9 respectively

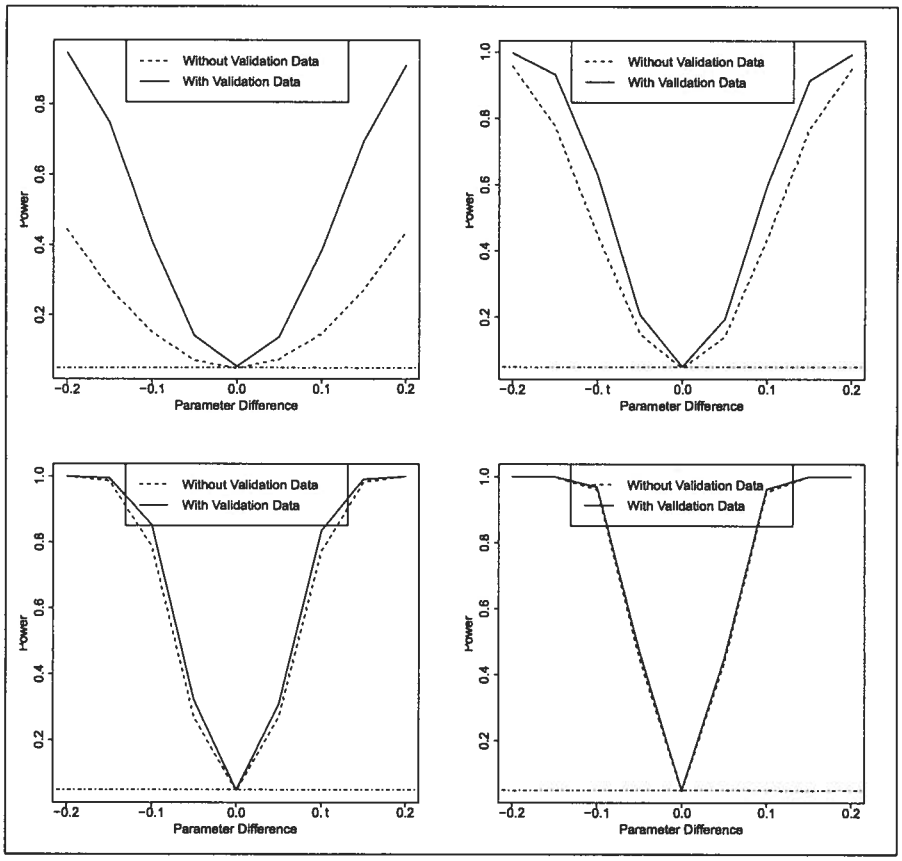
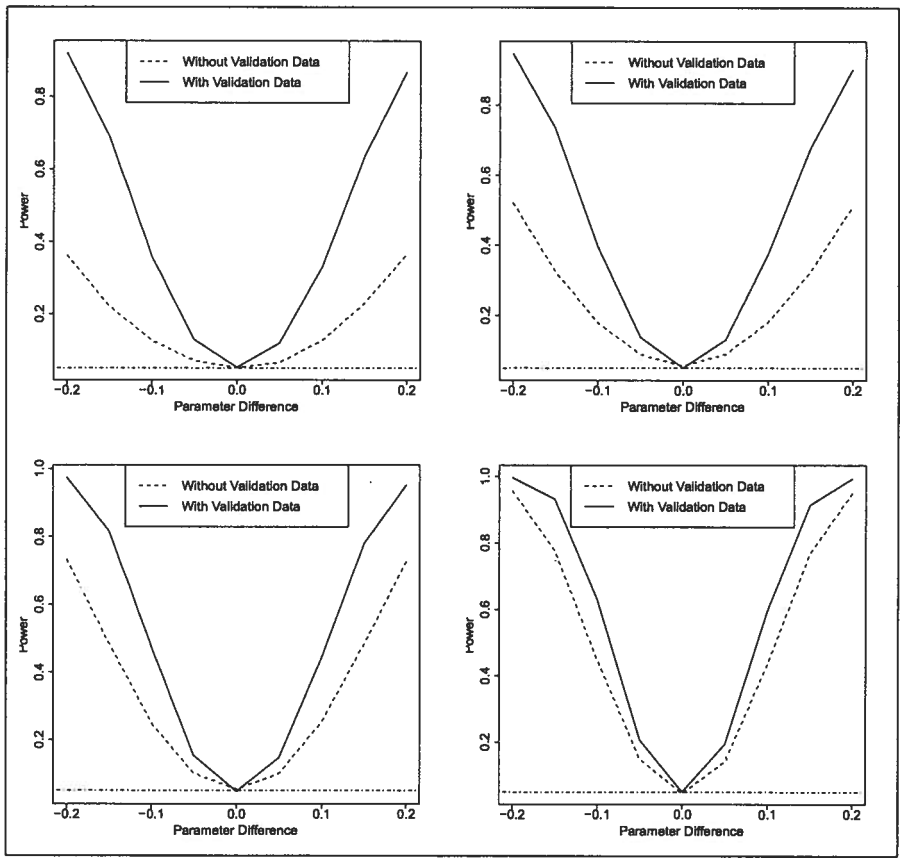


Figure 4.2: Power curves under different sample sizes: 400, 600, 1000 and 2000 respectively (validation sub-sample size is fixed at 200 in each situation)



which are shown in Figure 4.4. In all the case, we find the previous conclusion is still valid.

One point we should mention is that, for lower exposure prevalence such as 0.005, while finding the maximum likelihood estimators, sometimes the `optim` function goes out of bound. Therefore, for finding maximum likelihood estimators of 10,000 simulations in the null hypothesis situation ($r_0 = r_1 = r$), we had to iterate the process of generating new datasets 132,134 times for $r = 0.005$, 42,798 times for $r = 0.01$, 10,859 times for $r = 0.05$ and 10,047 times for $r = 0.10$. However, for higher exposure prevalence rates, we never had this problem of non-convergence.

4.2.4 Under Different Proportion of Validation and Main Part of the Data

As for all the other scenarios, we generated 10,000 datasets, with sensitivity and specificity 0.7 and exposure prevalence 0.4. But, keeping the total sample size fixed at 2000, we changed the proportions of the validation and the main (unvalidated) part of the dataset, – which are 1:9 (200:1800), 1:3 (500:1500), 1:1 (1000:1000) and 3:1 (1500:500) respectively. From Figure 4.5, the two parameter model has an identical power curve in all situations, but as the proportion of main data decreases for the four parameter model, power increases sharply.

4.2.5 Comparison under Budgetary Constraint

Cost effectiveness is obviously an important measure of the ultimate worth of a study design. While designing a study, we aim to obtain the best quality of information for a given resource, say, in terms of money or time. Of course, the optimal solution for a given study design is hard to obtain, because not all the parameters are usually known and there might be external constraints. Nonetheless, for our study, by considering the stated assumptions and the parameters of the described models, we tried to find which model performs better under a fixed budgetary constraint.

Validation data is costly to collect. The high cost of validation data limits the size of the validation sub-sample in a fixed cost design. From the previous scenarios we considered, the model without validation data could be at best as good as the model with validation data given favorable conditions, but never better. The critical issue we wanted to investigate here is to

Figure 4.3: Power curves under different Exposure Prevalences: 0.25, 0.30, 0.35 and 0.4 respectively

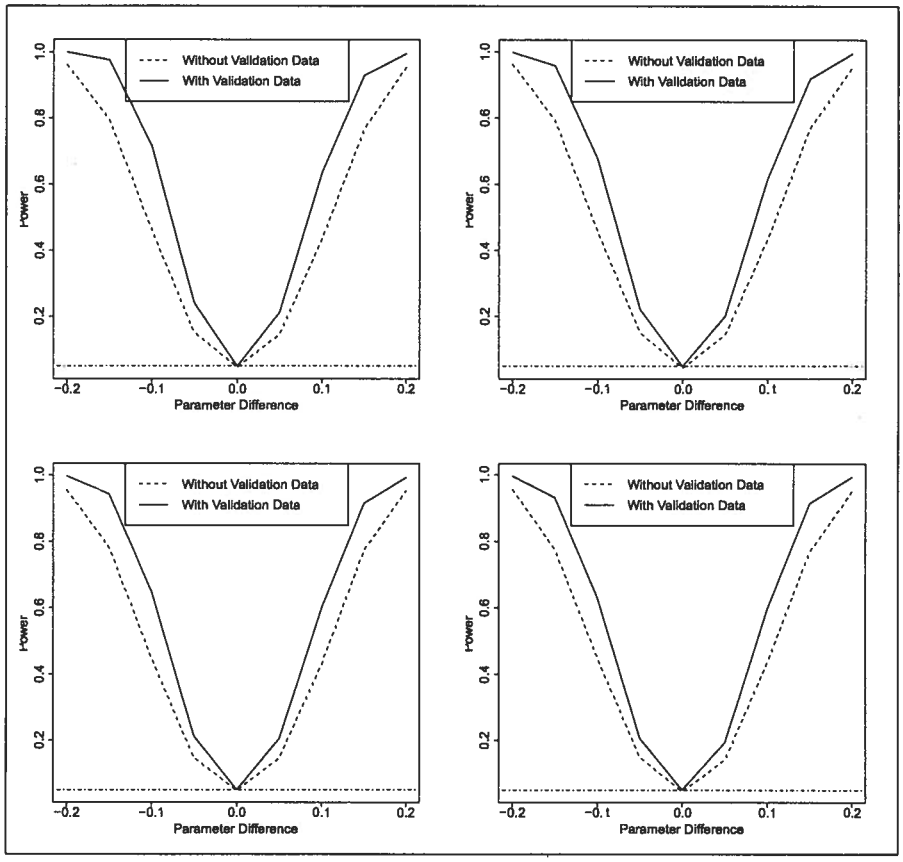


Figure 4.4: Power curves under smaller Exposure Prevalences: 0.005, 0.01, 0.05 and 0.10 respectively

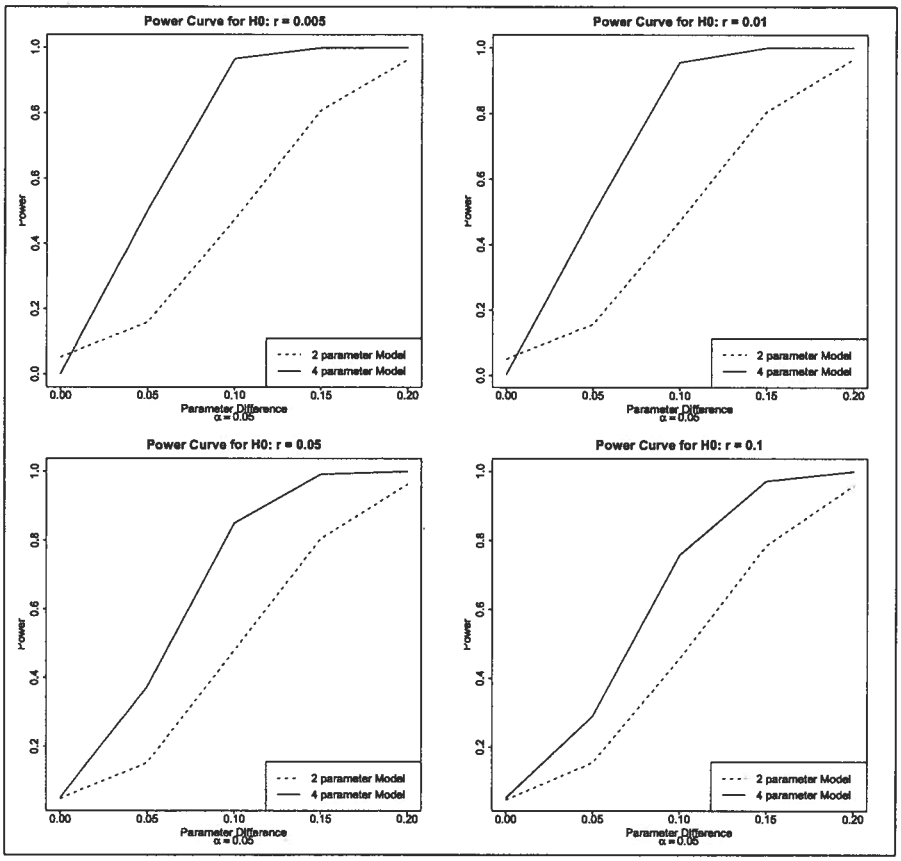
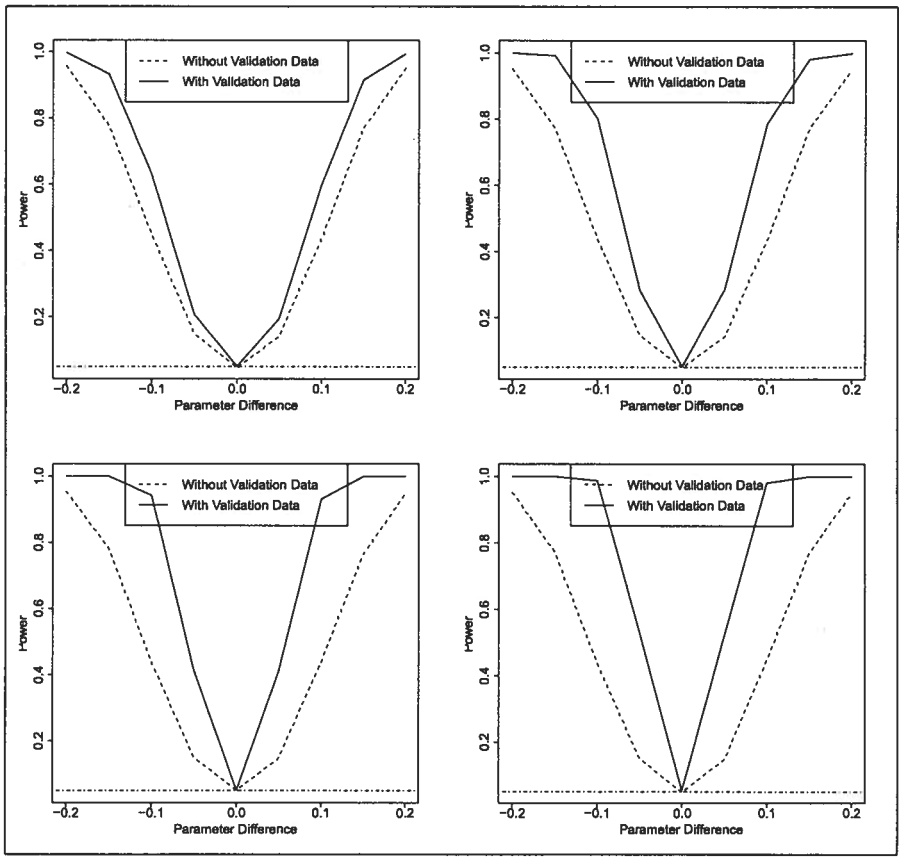


Figure 4.5: Power curves under different proportions of validation part and main part of the data: (1:9, 1:3, 1:1 and 3:1) respectively



find out whether there is any point where the model without validation data becomes superior to the model with validation data. In other words, how costly the validation data have to be to abandon the model with validation data, or whether a researcher can always choose the model with validation data without any trade-off. We investigated using a particular example as follows.

Say, we have \$2400 as a budget for designing a retrospective study using either the model with validation data or the model without validation data. We arbitrarily set \$1 as the cost of an unvalidated observation. We consider three pricing choices:

1. Collecting validated data costs three (3) times cost as much as collecting unvalidated (main) data. The allocations of validated and unvalidated data considered are provided in Table 4.2.

2. Collecting validated data costs five (5) times cost as much as collecting unvalidated (main) data. The allocation of validated and unvalidated data considered are provided in Table 4.3.

3. Collecting validated data costs ten (10) times cost as much as collecting unvalidated (main) data. The allocations of amount of validated and unvalidated data considered are provided in Table 4.4.

Table 4.2: Scenarios under constant cost = \$2400, assuming that collecting validated data costs three (3) times as much as collecting unvalidated (main) data

Scenario	Validated data	Unvalidated data	Cost
Q.3	2×50	2×1050	$2 \times (3 \times 50 + 1050) = 2400$
R.3	2×100	2×900	$2 \times (3 \times 100 + 900) = 2400$
S.3	2×200	2×600	$2 \times (3 \times 200 + 600) = 2400$
T.3	2×300	2×300	$2 \times (3 \times 300 + 300) = 2400$

In Tables 4.2, 4.3 and 4.4, we only consider situations where sample sizes are equal for cases and controls in both validated and unvalidated parts.

From Figure 4.6, the with validation data model is still superior in all scenarios despite the fact that validation sample costs three times more to collect compared to an unvalidated sample. However, when the cost is five times as much, both models have almost the same utility, as shown in Figure 4.7. But from Figure 4.8, it is evident that the model without validation

Figure 4.6: Power curves under fixed amount of cost = \$2400 assuming that collecting validated data costs three (3) times as much as collecting unvalidated (main) data

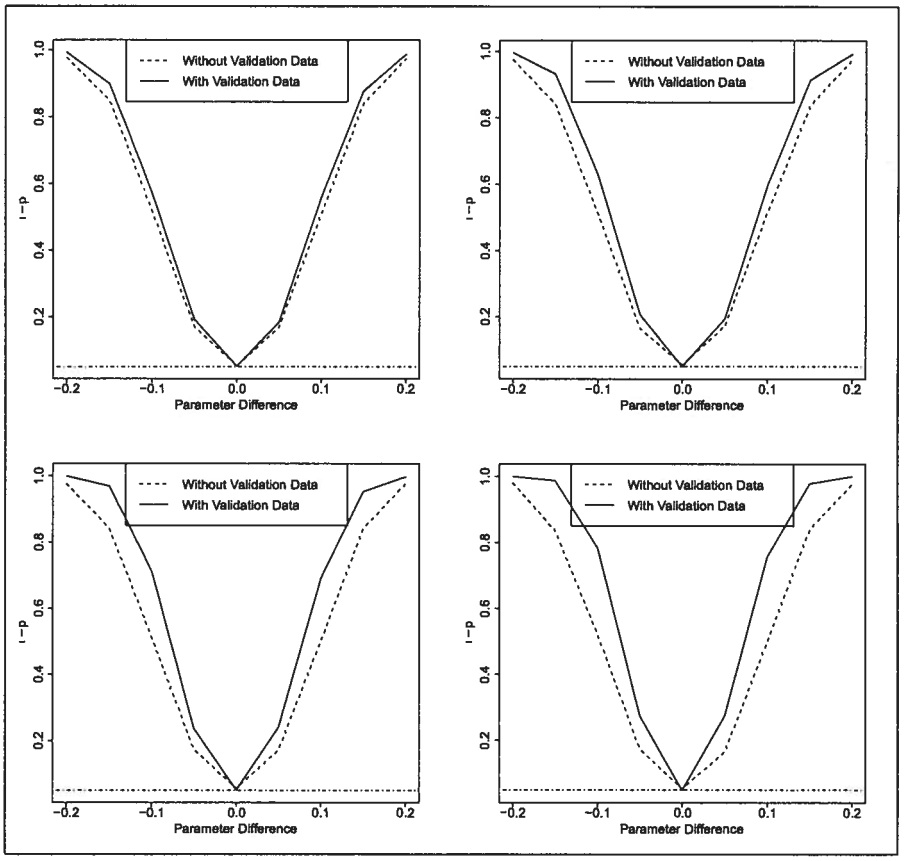


Table 4.3: Scenarios under constant cost = \$2400, assuming that collecting validated data costs five (5) times as much as collecting unvalidated (main) data

Scenario	Validated data	Unvalidated data	Cost
Q.5	2×50	2×950	$2 \times (5 \times 50 + 950) = 2400$
R.5	2×100	2×700	$2 \times (5 \times 100 + 700) = 2400$
S.5	2×150	2×450	$2 \times (5 \times 150 + 450) = 2400$
T.5	2×200	2×200	$2 \times (5 \times 200 + 200) = 2400$

Figure 4.7: Power curves under fixed amount of cost = \$2400 assuming that collecting validated data costs five (5) times as much as collecting unvalidated (main) data

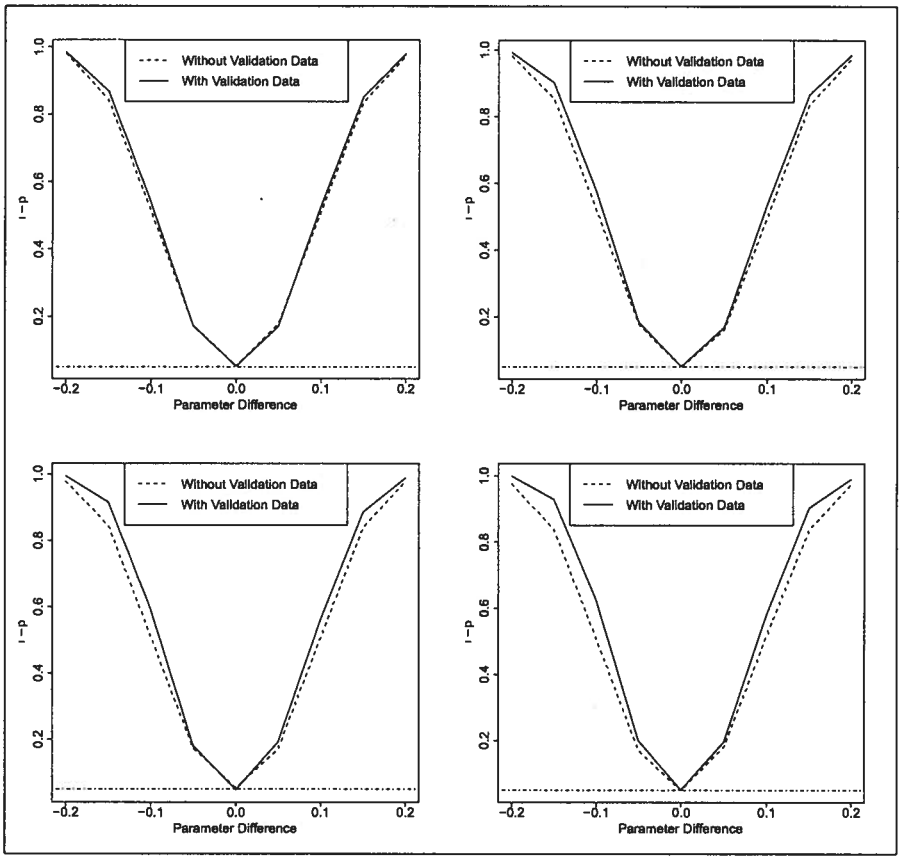
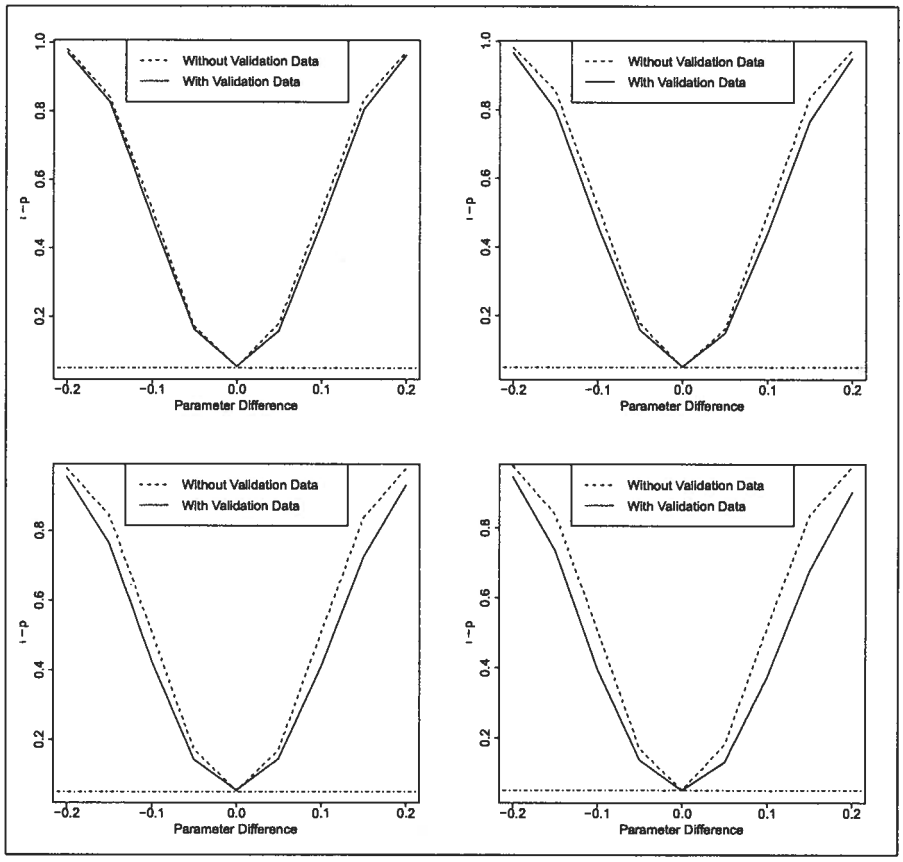


Table 4.4: Scenarios under constant cost = \$2400, assuming that collecting validated data costs ten (10) times as much as collecting unvalidated (main) data

Scenario	Validated data	Unvalidated data	Cost
Q.10	2×25	2×950	$2 \times (10 \times 25 + 950) = 2400$
R.10	2×50	2×700	$2 \times (10 \times 50 + 700) = 2400$
S.10	2×75	2×450	$2 \times (10 \times 75 + 450) = 2400$
T.10	2×100	2×200	$2 \times (10 \times 100 + 200) = 2400$

Figure 4.8: Power curves under fixed amount of cost = \$2400 assuming that collecting validated data costs ten (10) times as much as collecting unvalidated (main) data



data can be superior to the with validation data model, given a fixed total cost and a much higher cost for validation data. This is the only case among the scenarios we have considered, where the model without validation data can possibly be superior - when the cost of a validated observation is much higher than the cost of an unvalidated observation. This is one practical limitation of the model with validation data that the researchers should keep in mind when designing a study.

4.3 Scenario Settings Under Bayesian Adjustment

While dealing with Bayesian adjustments, we utilize all the n's (observed values) and the u's (unobserved values) in Tables 1.1 and 1.2, and the models (two parameter for data without validation part and four parameter models for data with validation part) discussed in §3.2.3 are utilized. As for frequentist adjustment, for the two parameter model, we simply use the column totals from the tables, while in the four parameter model, we use the n's that are inside the validation table as they are observed. To ensure comparability, both models in Bayesian adjustment utilize the same amount of data. The only difference is that the four parameter model recognizes the validation part, while the two parameter model ignores the true classification information of the validation part.

Exactly the same scenarios discussed in §4.2 are considered to understand the performance of Bayesian adjustment to nondifferential misclassification, varying the level of exposure prevalence, or sensitivity and specificity, or sample size, or sample proportion of the validation and the main (unvalidated) part of the data.

We used the power curve from the likelihood ratio tests as the comparison tool in assessing the frequentist adjustments. However, finding such a tool for Bayesian adjustment models was not straightforward. Instead, this is what we have done: Once we have generated the data (as discussed in §4.1), we implemented the mixed algorithm as described in §3.2.3 for 10,000 Markov Chain Monte Carlo iterations. Half of the Markov Chain Monte Carlo iterations were discarded as burn-in (we will justify the length of chain and burn-in in §4.3.5). Using the retained chains, we constructed a 95% credible interval for the odds-ratio and checked whether this credible interval contained the null value ($OR = 1$) or not (where OR is a function of r_0 and r_1 for four parameter model as given in Equation (1.2), and also, for two parameter model, OR is a function of θ_0 and θ_1 as given in Equation (1.3)). One other way of serving this same purpose would be to construct 95% credible interval for the logarithmic transformation of the odds-ratio and test whether the constructed credible interval contains the null value ($\log OR = 0$) or not. 2,000 datasets for each set of parameters in a particular case of the scenario. To produce a graph for the cases of each scenario, we do it for nine points (corresponding to the differences of the alternative

hypothesis, as was discussed in §4.2 for the frequentist approach of power curve construction procedure) for each of the models. This information of what proportion of credible intervals excluded the null value was used to find a probabilistic solution of the problem of comparison. This tool could also be labeled as a kind of power curve since this also uses the similar theme "reject H_0 if the credible interval excludes null value", instead of the statement "reject H_0 if the p-value is less than significance level". From the deviation from one model's curve to another, one can have some understanding of the performance of the two models in these situations.

Again, to allow reproducibility of the results, the seed is chosen arbitrarily and kept the same throughout the entire analysis. Initial values need to be provided for Markov Chain Monte Carlo algorithms. Experience suggests that the initial values does not have much impact on the final results. Details of this comment are shown in §4.3.5.

4.3.1 Under Different Values of Sensitivity and Specificity

The same cases as considered in frequentist adjustment are carried out. From the Figure 4.9, it is evident that the two parameter model is always dominated by the four parameter model.

4.3.2 Under Different Sample Sizes

Figure 4.10 shows that the tests get better as the sample size increases, but the four parameter model is always better than the two parameter model.

4.3.3 Under Different Exposure Prevalence Rates

The graphs under consideration are shown in Figure 4.11. In all the cases, two parameter models are dominated by the four parameter models, and the graphs of credible intervals excluding null values do not seem to vary much for the various prevalence rates under consideration.

4.3.4 Under Different Proportion of the Validation Data

From Figure 4.12, the two parameter model in all cases have almost similar curves, but the powers increase sharply as the proportion for the validation part increases for four parameter model.

Figure 4.9: Bayesian analysis results for different sensitivity and specificity values (.6, .7, .8, .9): Proportion of credible intervals excluding null value

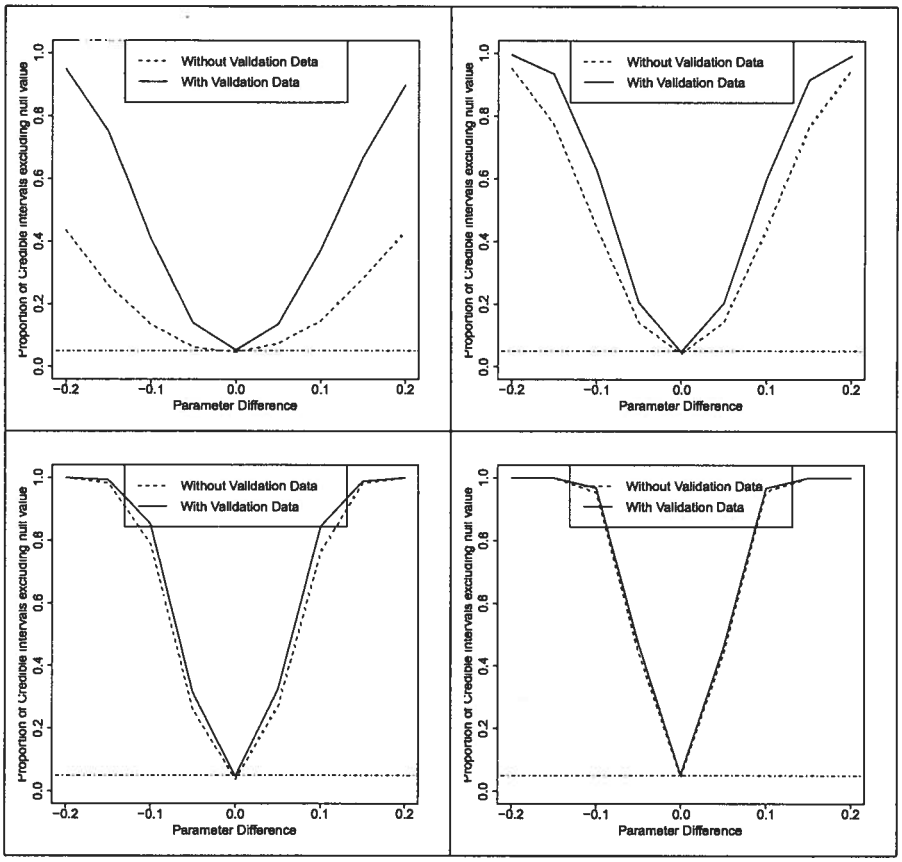


Figure 4.10: Bayesian analysis results for different sample sizes (400, 600, 1000, 2000): Proportion of credible intervals excluding null value

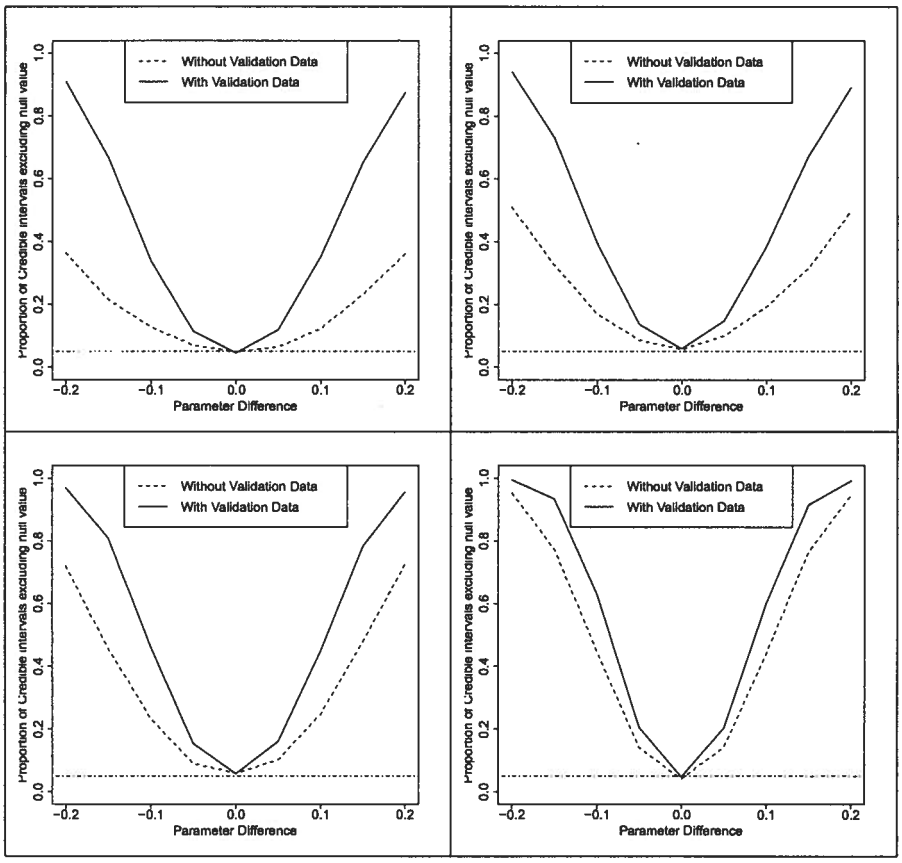


Figure 4.11: Bayesian analysis results for different exposure prevalence (.25, .3, .35, .4): Proportion of credible intervals excluding null value

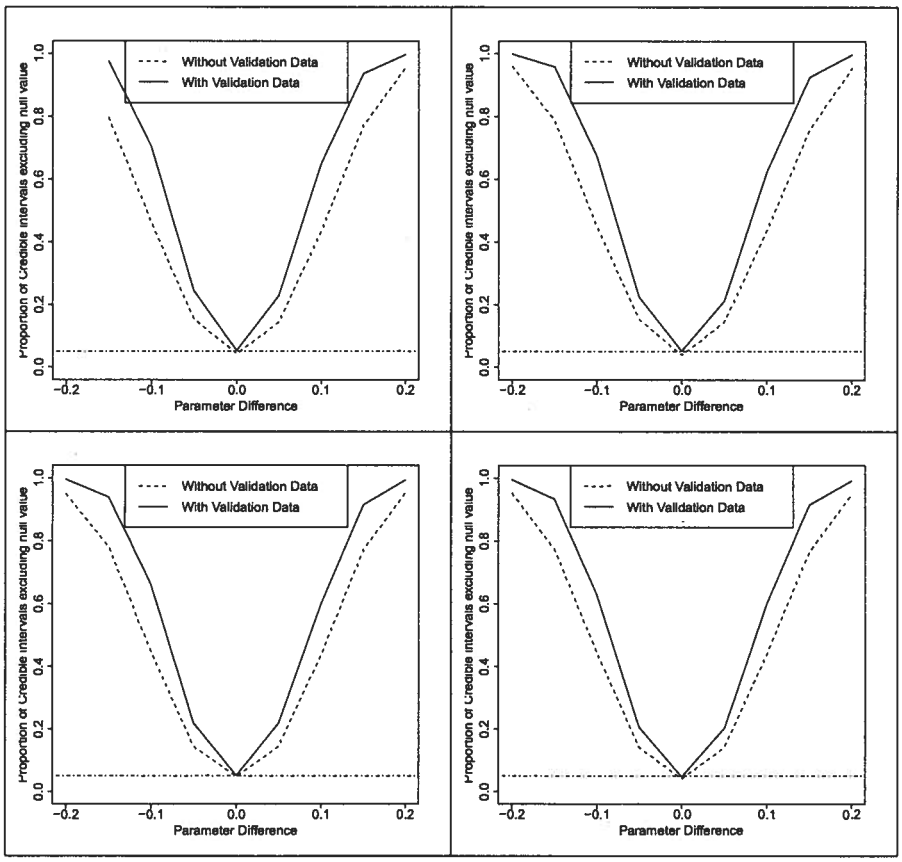
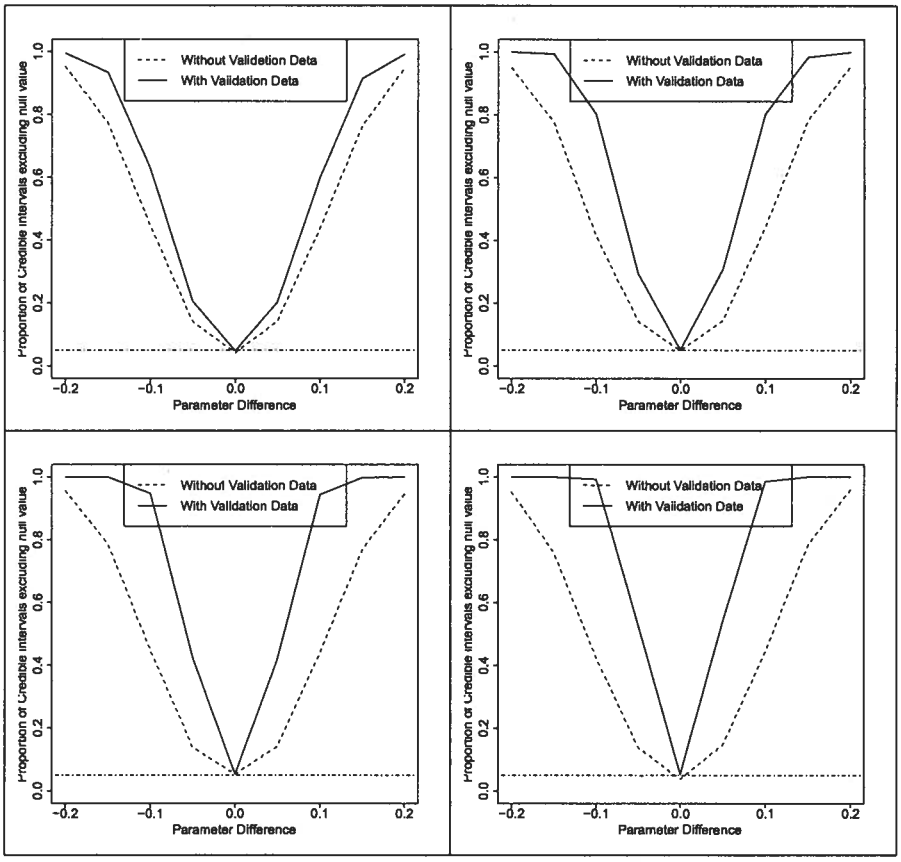


Figure 4.12: Bayesian analysis results for different ratio of data splits (1:9, 1:3, 1:1, 3:1 respectively for validation and main part): Proportion of credible intervals excluding null value



In all four situations considered here, we have the same conclusion about the respective situations from both the frequentist and Bayesian approaches. In fact, if we compare Figure 4.9 with Figure 4.1 and Figure 4.10 with Figure 4.2 and Figure 4.11 with Figure 4.3 and Figure 4.12 with Figure 4.5, the shapes of curves from the respective situations are strikingly similarity.

4.3.5 Diagnostics

For diagnostic purposes, we generate datasets with exposure prevalence 0.3 for both case and control groups and sensitivity and specificity both equals to 0.7. As shown in Figures 4.13, 4.14, 4.15 and 4.16 for 10,000 MCMC iterations, the trace plot of all the parameters r_0, r_1, SN and SP look stable after the burn-in in four chains with different starting values (0.2, 0.4, 0.6, 0.8 for each parameters). All these figures are obtained using one single dataset as an example. The burn-in is colored as grey and after burn-in, the estimates are colored as black in each of these graphs.

Sometimes graphical diagnostics are not very reliable. Therefore, we resort to some statistics that are used for such chain diagnosis, such as Gelman and Rubin's convergence diagnostic statistic which was discussed in §3.3.2. This statistic requires more than one chain, and hence we used the four chains with four different set of initial values. Theory says that the statistic should not go beyond 1.2. For this particular dataset, for the parameters r_0, r_1, SN and SP, we had the Gelman and Rubin's convergence diagnostic statistic, $\hat{R}$ = 1.011, 1.003, 1.011 and 1.008 respectively, for 10,000 iterations in each. Figure 4.21 indicates the evolution of Gelman and Rubin's convergence diagnostic statistic as the number of iterations increases. From this figure, it is evident that the chain is very satisfactory after burn-in.

Also, to check whether the initial value has any effect on estimates, we plotted the means of each of the four chains which started from different initial values. As shown in Figures 4.17, 4.18, we can see that both converge to 0.3, which was the original exposure prevalence value used to generate the considered dataset. Similarly, from Figures 4.19, 4.20, we see that both the sensitivity and specificity eventually converge to 0.7, which was the parameter value used to generate the datasets under consideration.

Figure 4.13: Diagnosis of convergence of Bayesian analysis results: Trace Plots for r_0 in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset

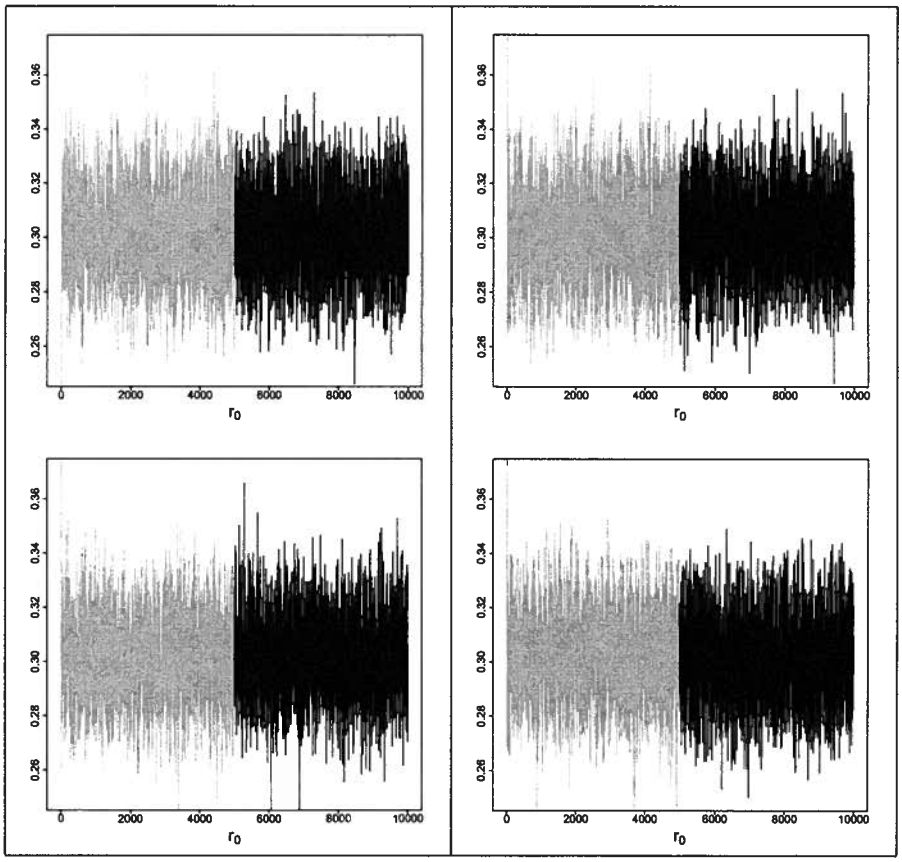


Figure 4.14: Diagnosis of convergence of Bayesian analysis results: Trace Plots for r_1 in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset

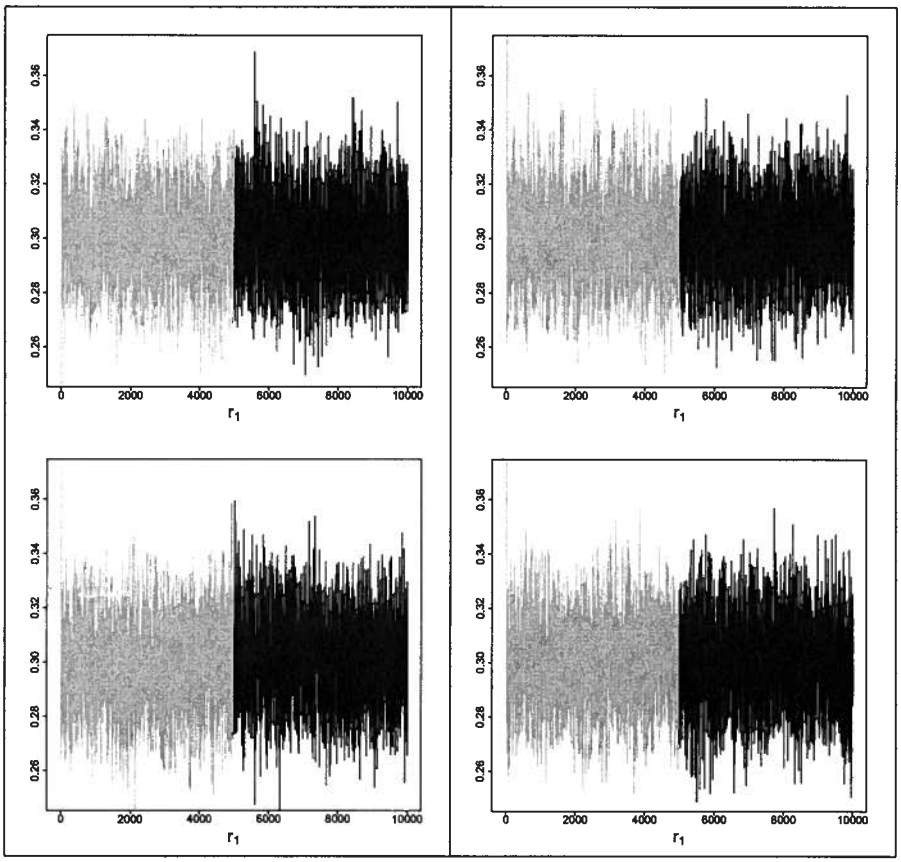


Figure 4.15: Diagnosis of convergence of Bayesian analysis results: Trace Plots for SN in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset

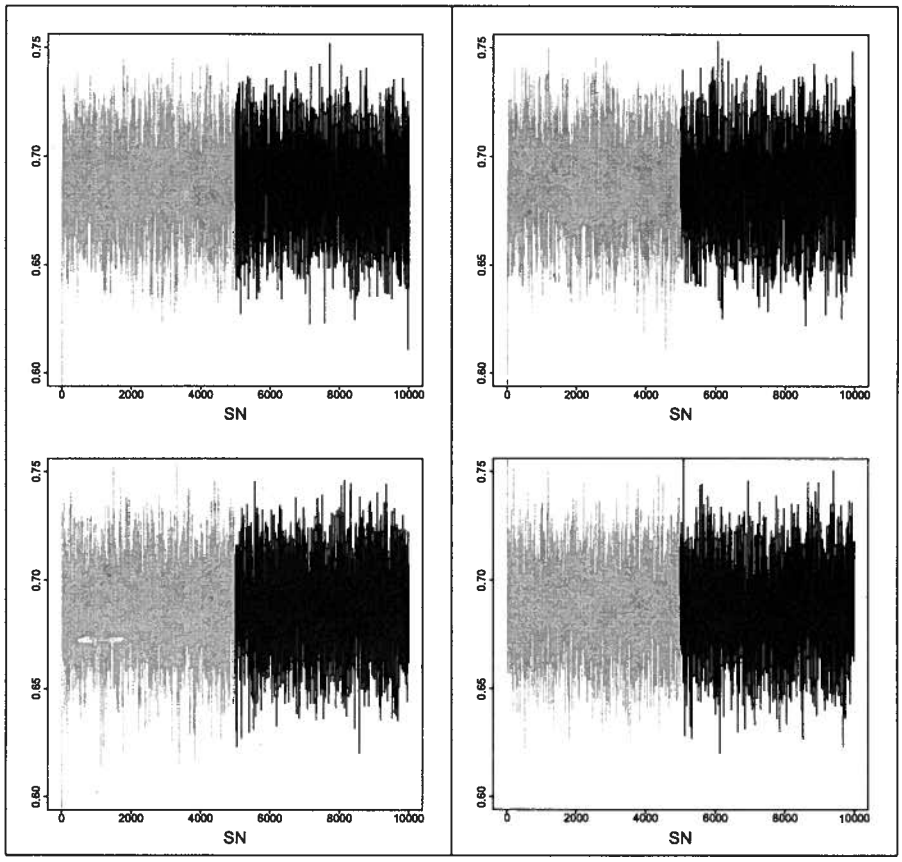


Figure 4.16: Diagnosis of convergence of Bayesian analysis results: Trace Plots for SP in 4 chains with different starting values (for 10,000 iterations, with half burn-in) for a single dataset

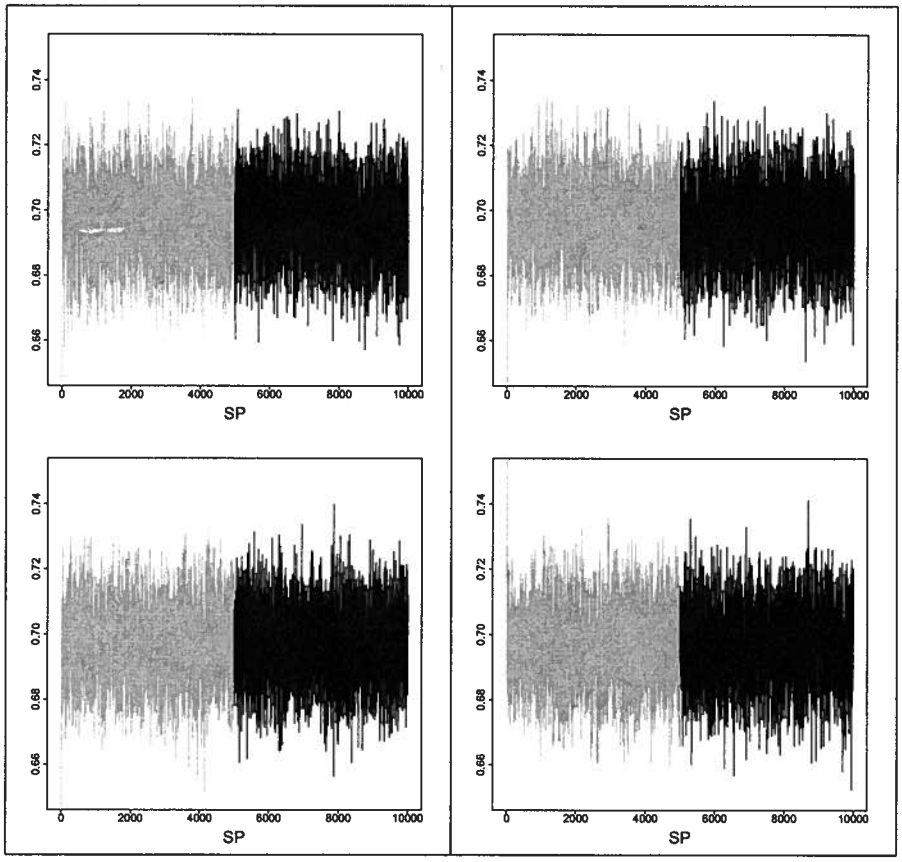


Figure 4.17: Sequence of the mean of posterior for r_0 for the four Markov Chain Monte Carlo Chains for 10,000 iterations

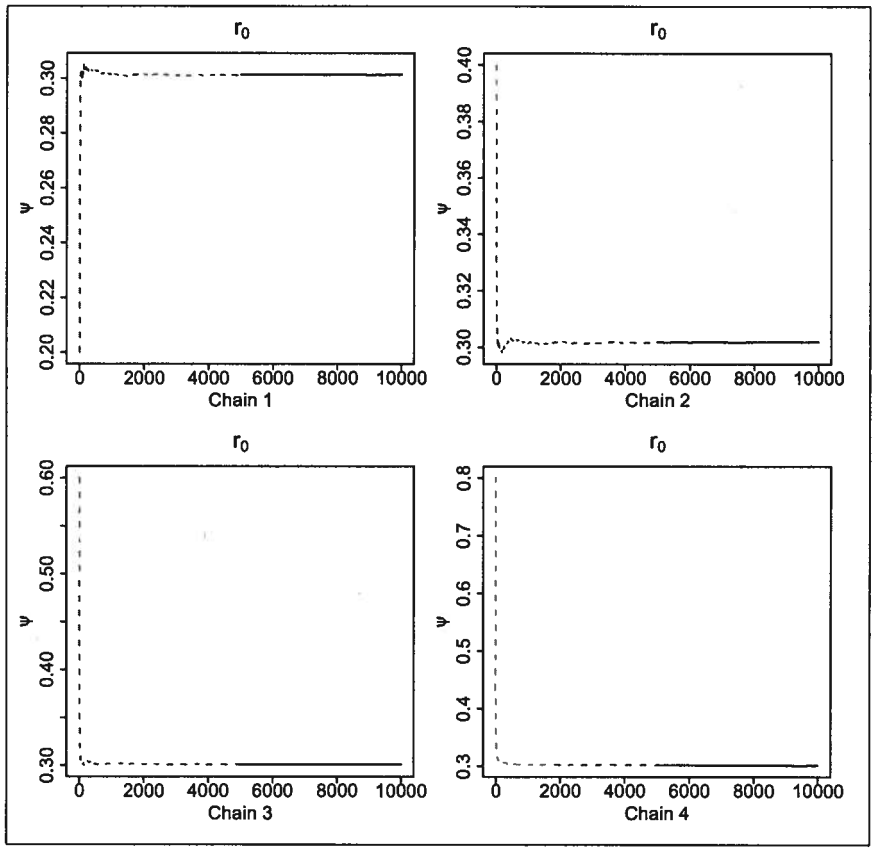


Figure 4.18: Sequence of the mean of posterior for r_1 for the four Markov Chain Monte Carlo Chains for 10,000 iterations

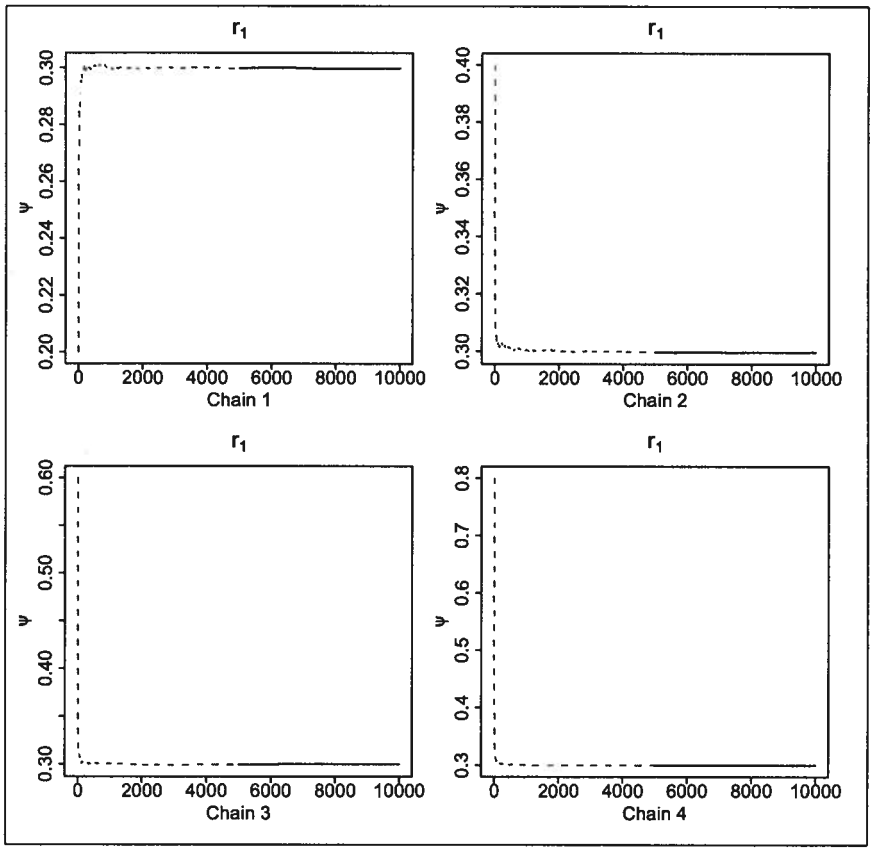


Figure 4.19: Sequence of the mean of posterior for SN for the four Markov Chain Monte Carlo Chains for 10,000 iterations

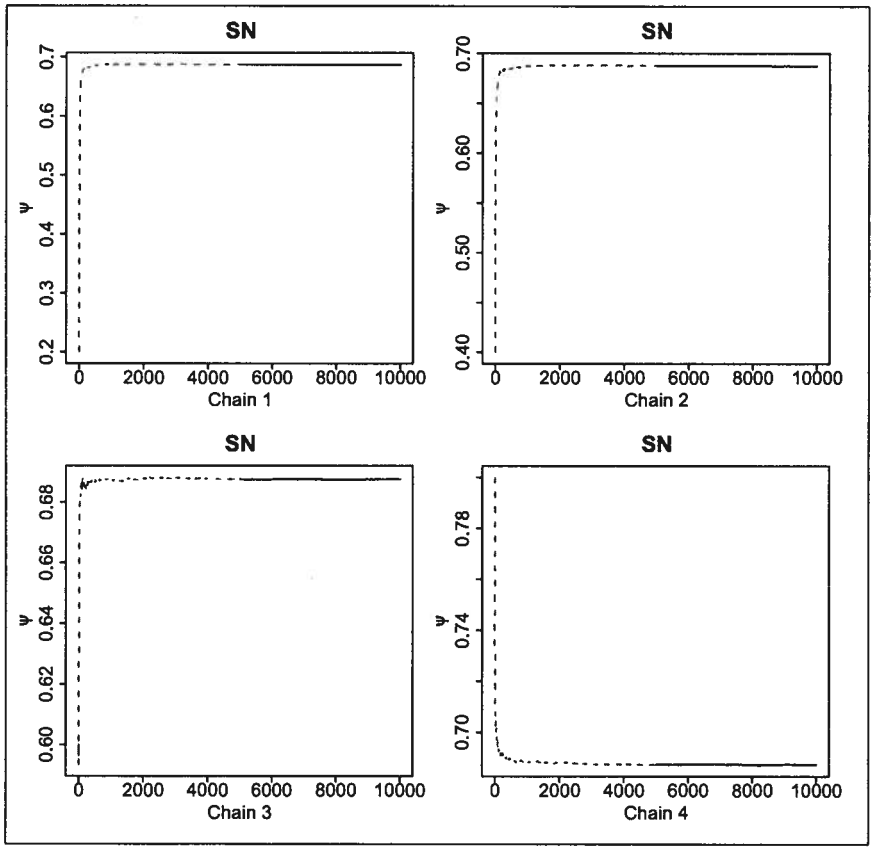


Figure 4.20: Sequence of the mean of posterior for SP for the four Markov Chain Monte Carlo Chains for 10,000 iterations

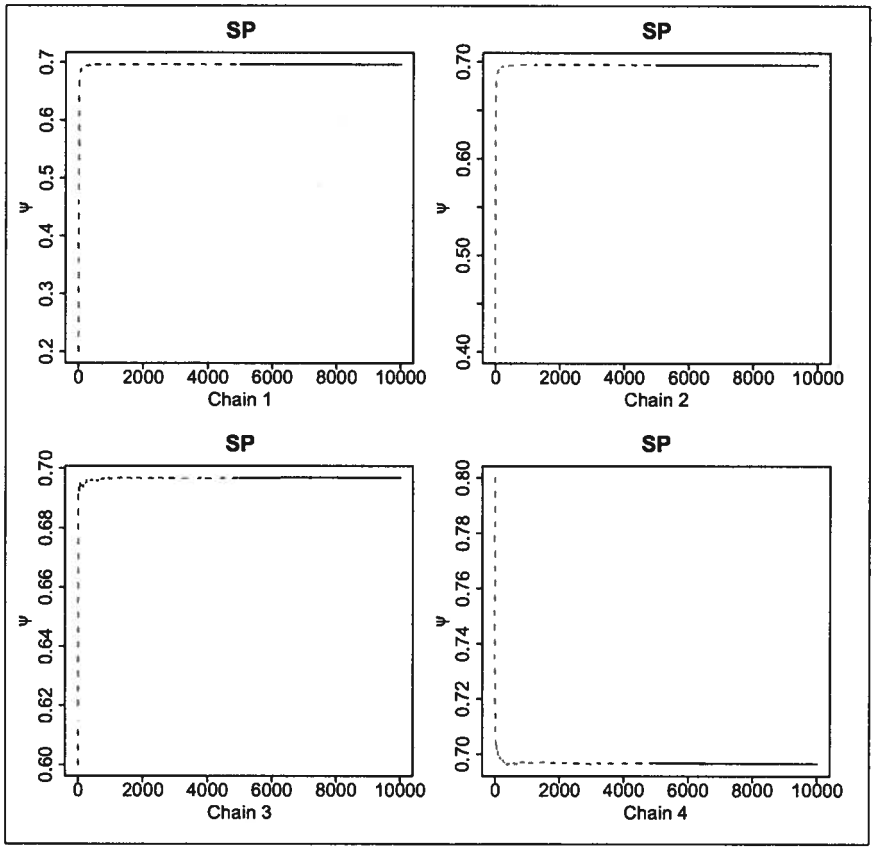
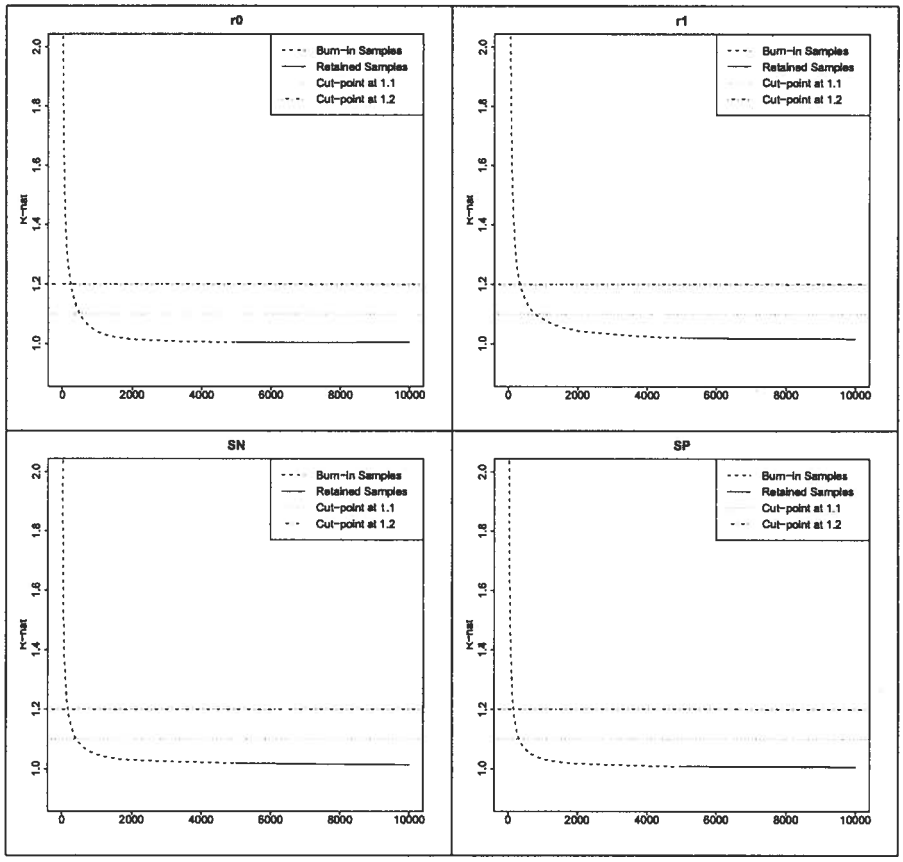


Figure 4.21: Sequence of the Gelman-Rubin $\hat{R}$ for the four Markov Chain Monte Carlo Chains for 10,000 iterations



Chapter 5

Application in Epidemiological Studies

5.1 Introduction

In almost all epidemiological studies, some amount of error in assessment is inevitable. The extent of such error depends on various factors, such as the nature of the exposure, and the instrument error associated with collecting the information. In this chapter, we will consider two epidemiologic datasets where challenges specifically arise in accurately identifying the outcome of exposure. The methods discussed in the previous chapters are considered and applied to these datasets.

5.2 Study of Sudden Infant Death Syndrome (SIDS)

The performance of the methods described in the previous chapters are illustrated using a case-control study of antibiotic prescription during pregnancy and subsequent occurrence of Sudden Infant Death Syndrome (SIDS) [*Greenland*, 1988b, a, 2008], [*Marshall*, 1990, 1997], [*Kraus et al.*, 1989]. The association of interest is between the prescription of antibiotics during pregnancy (V) and SIDS (Y). The surrogate exposure or error-prone measurement (V^*) was an interview response, whereas the true exposure (V) was derived from medical records. The validation studies, in cases ($Y = 1$) and controls ($Y = 0$), were joint (V^*, V) designs done as sub-studies that resulted in the data presented in Table 5.1.

Frequentist estimates of parameters of the two models under consideration for the SIDS study data are reported in Table 5.2. From this table, we can see that the apparent prevalence rates are close estimates of the prevalence rates obtained while considering the validation data. The validation data model shows that the data has low sensitivity (0.6), but high speci-

Table 5.1: Data from the study of sudden infant death syndrome (SIDS) and antibiotic prescription

Y	Cases ($Y = 1$)		Controls ($Y = 0$)	
Validated Part	$V^* = 1$	$V^* = 0$	$V^* = 1$	$V^* = 0$
$V = 1$	29	17	21	16
$V = 0$	22	143	12	168
Unvalidated (main)	122	442	101	479
Total	173	602	134	663

ficity (0.9). The log-odds ratios in both groups are positive numbers. For without validation data, the estimate of odds ratio is 1.422 and 95% Wald confidence limits are (1.11, 1.83), calculated using the formula provided by *Marshall* [1997]. Not surprisingly, the likelihood ratio p-value obtained from this model is small (0.006). These results match with the case discussed by *Greenland* [2008]. On the other hand, for with validation data model, the estimated odds ratio is 1.49 with 95% Wald confidence limits (1.02, 2.16), which is coherent with the findings of *Greenland and Gustafson* [2006]. The likelihood ratio p-value is also small in this model (0.035). Therefore, the conclusions from both models are the same. They suggest that the hypothesis $H_0 : r_0 = r_1$ is rejected at $\alpha = 0.05$. That is, the true $\log(OR)$ is significantly far away from 0 based on the evidence provided by the SIDS study data.

Table 5.2: Frequentist Estimates of the model parameters in the SIDS study

Not considering Validation data			Considering Validation data		
Parameters	Estimate	S.E.	Parameters	Estimate	S.E.
θ_0	0.168	0.013	r_0	0.163	0.021
θ_1	0.223	0.015	r_1	0.225	0.024
			SN	0.603	0.047
			SP	0.903	0.013
$\log(OR)$	0.352	0.128	$\log(OR)$	0.398	0.191
P-value	0.006		P-value	0.035	

The Bayesian estimates and standard errors are reported in Table 5.3. The priors used here are very general and similar to those described in §3.2.3. These results are very similar to those obtained using maximum likelihood.

Both the 95% credible intervals of the odds ratio obtained from the with and without validation data model fail to include the null value 1 inside the interval. Moreover, the estimates and credible intervals are very similar to those obtained by frequentist methods. Therefore, the null hypothesis is still rejected by the Bayesian tools. That means the data suggests a positive association between the prescription of antibiotic and consequent incidence of SIDS, under the assumption of equality of misclassification probabilities.

For the Bayesian estimates and hypothesis testing results reported in Table 5.3 and trace plots in Figure 5.1, the initial values of r_0, r_1, SN and SP were set to 0.4, 0.4, 0.7 and 0.7 respectively. For θ_0 and θ_1, it was 0.2 and 0.2.

One interesting issue needs to be addressed here. Other than a few spacial cases, it is well known that under the nondifferential misclassification assumption, in absence of any other errors, the estimates of measure of association, such as odds ratio should be biased towards the null 'on average'. However, in this particular data, we notice that the estimate of odds ratio slightly goes away from null (1.42 to 1.49), as the theory suggests, but the p-value from Wald test gives us the opposite message - it increases from 0.0029 to 0.0184 respectively (which dictates towards the null behavior after adjustment). This might be due to the fact that the posterior variance is being underestimated in the without validation data situation, and hence the posterior variance increases after adjustment, providing an even wider credible interval. Such phenomenon of increment of uncertainty even though the odds ratio moves away from null after adjustment, is already noted by *Gustafson and Greenland* [2006]. The likelihood ratio test acts similarly to the approximate Wald test. This might be one indication that the assumption of nondifferentiality was not completely satisfied, if we rule out the explanation of random variation due to chance in this particular example. Since both p-values are small enough to reject the null hypothesis, this does not alter the conclusion in this example.

The Gelman and Rubin convergence diagnostic statistic, $\hat{R}$ value for the four parameters r_0, r_1, SN and SP are 1.002, 1.003, 1.045 and 1.009 respectively. Also, for θ_0 and θ_0, $\hat{R}$ gives 1.002 and 1.003 respectively. All these values are much less than 1.2. Here, various initial values were set to check the convergence - such as 0.2, 0.4, 0.6 and 0.8 for each of the parameters under consideration. 10,000 iterations were performed and half were retained after burn-in to estimate each parameters. Also, from Figure 5.2, we can see that the posterior distributions does not have any multimodality, which

Figure 5.1: MCMC for the with and without validation data model parameters in the SIDS study

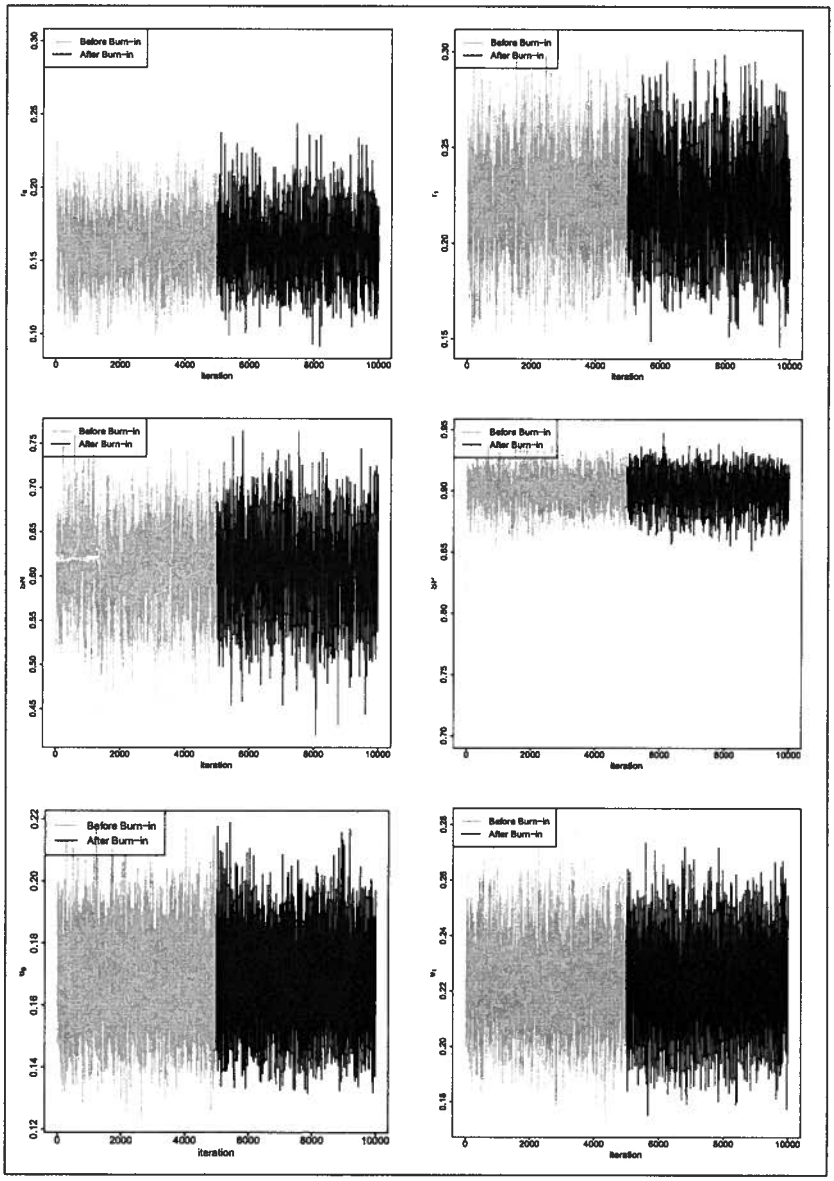


Table 5.3: Bayesian Estimates of the model parameters in the SIDS study

Not considering Validation setting			Considering Validation setting		
Parameters	Estimate	SD	Parameters	Estimate	SD
θ_0	0.168	0.013	r_0	0.161	0.020
θ_1	0.222	0.015	r_1	0.221	0.024
			SN	0.609	0.046
			SP	0.901	0.012
$\log(OR)$	0.351	0.129	$\log(OR)$	0.395	0.186
95%C.I. (OR)	Does not include H_0 value (1.103, 1.830)		95%C.I. (OR)	Does not include H_0 value (1.038, 2.153)	

is a sign of good convergence.

5.3 Cervical Cancer and Herpes Simplex Virus Study

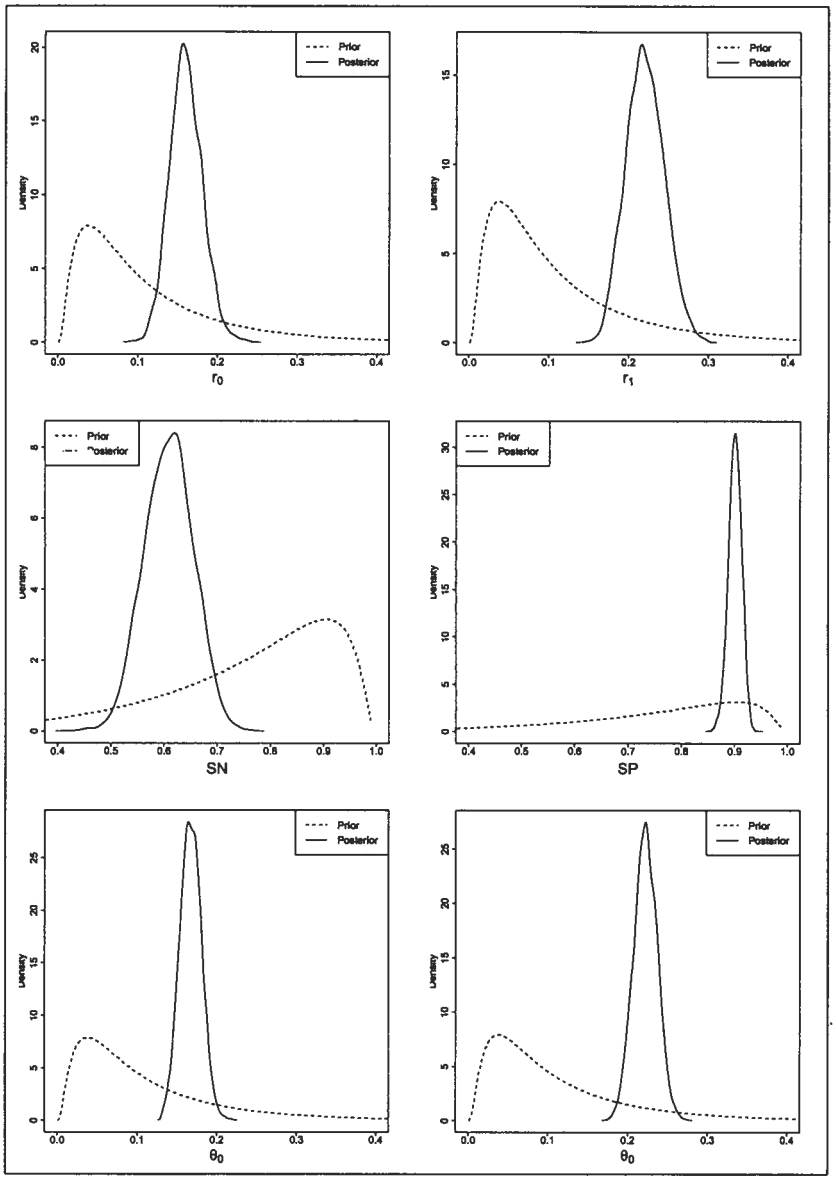
This data is listed in *Carroll et al.* [1993] and discussed in *Prescott and Garthwaite* [2002], *Carroll et al.* [2006]. The research question is whether exposure to herpes simplex virus contributes to the risk of cervical cancer. The response variable Y is an indicator of cervical cancer, V is exposure to type 2 herpes simplex virus (HSV-2) measured by a refined western blot procedure and V^* is exposure to HSV-2 measured by the western blot procedure. The data is provided in Table 5.4.

Table 5.4: Data from Herpes Simplex Virus-2 study

Y	Cases ($Y = 1$)		Controls ($Y = 0$)	
Validated Part	$V^* = 1$	$V^* = 0$	$V^* = 1$	$V^* = 0$
$V = 1$	18	5	16	16
$V = 0$	3	13	11	33
Unvalidated (main)	375	318	535	701
Total	396	336	562	750

Frequentist estimates of parameters of the two models under consideration for the HSV-2 study data are reported in Table 5.5. From this table, we can see that the apparent prevalence rates and estimates of the prevalence

Figure 5.2: Prior and Posterior Distributions of all the Parameters under Consideration in the SIDS study



rates for the case group obtained in the presence of the validation data are not in complete agreement. Especially the prevalence rates for case group are much higher than the control group, both in the before and after adjustments. For without validation data, the estimated odds ratio is 1.57 and 95% Wald confidence limits are (1.31, 1.89). Also, the likelihood ratio p-value obtained from this model is very small. On the other hand, for the with validation data model, the estimated odds ratio is 2.61 with 95% Wald confidence limits (1.62, 4.18). The likelihood ratio p-value is also very small in this model. Since the p-values obtained from both models are very small, the conclusions from both models are the same. They suggest that the hypothesis $H_0 : r_0 = r_1$ is rejected at $\alpha = 0.05$. The validation data model shows that the exposure assessment has moderate sensitivity, as well as moderate specificity.

Table 5.5: Frequentist Estimates of the model parameters in the HSV-2 study

Not considering Validation setting			Considering Validation setting		
Parameters	Estimate	S.E.	Parameters	Estimate	S.E.
θ_0	0.428	0.014	r_0	0.418	0.046
θ_1	0.541	0.018	r_1	0.652	0.053
			SN	0.679	0.041
			SP	0.743	0.043
$\log(OR)$	0.453	0.093	$\log(OR)$	0.958	0.237
P-value	9.966×10^{-7}		P-value	1.48×10^{-6}	

The Bayesian estimates and standard errors are reported in Table 5.6. For the model without validation data, these results are almost the same as those obtained using maximum likelihood. However, the estimates obtained from the model with validation data are not nearly as close. Nonetheless, both the 95% credible intervals of the odds ratio obtained from the with and without the validation data model fail to include the null value 1 inside the interval. Therefore, even with the Bayesian method, the null hypothesis is rejected. Moreover, the conclusions obtained from hypothesis testing are very similar to those obtained by frequentist methods. As a result, we can conclude that the exposure of HSV-2 is positively associated with increased risk of developing cervical cancer.

Using the same prior that we used in simulations in chapter 4, we can see that the exposure prevalences are greatly underestimated in prior den-

sities. The posterior exposure prevalences are very different than suggested in the prior. From Figure 5.4 it is evident that the posterior results are not dominated by the given prior.

Table 5.6: Bayesian Estimates of the model parameters in the HSV-2 study

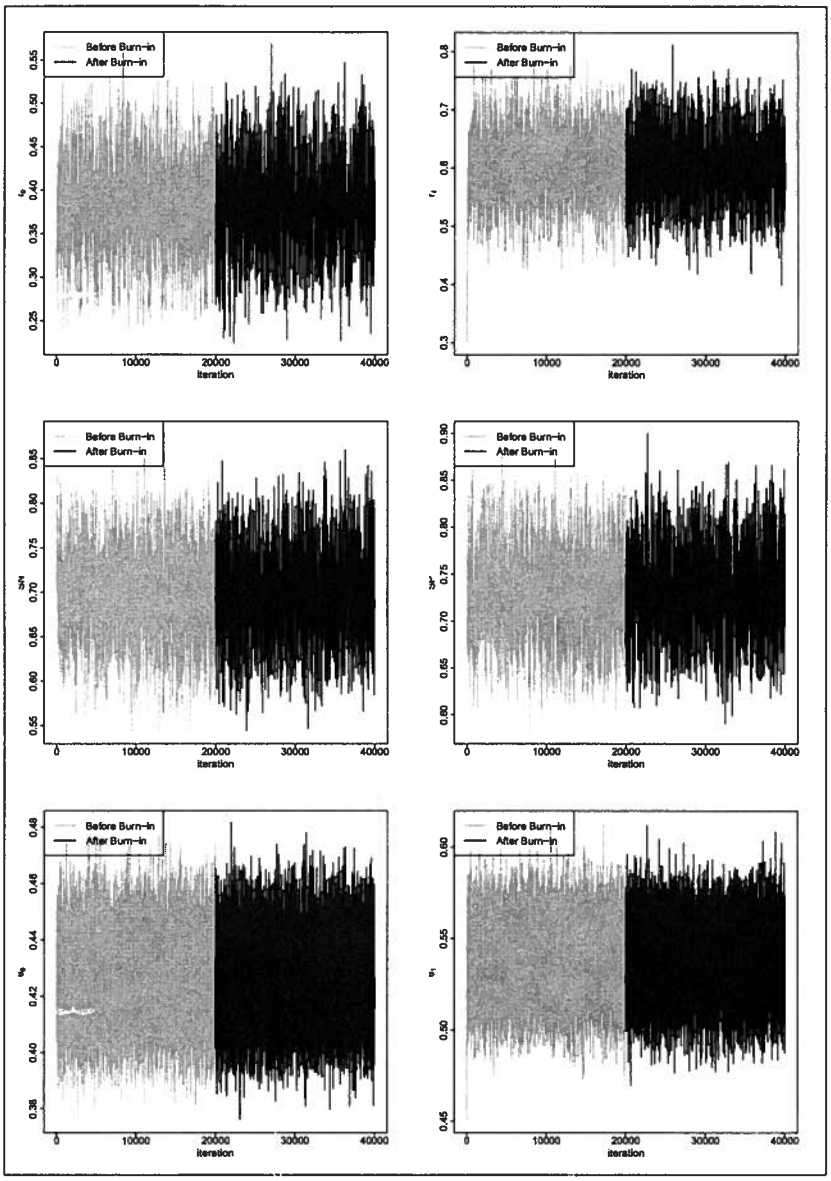
Not considering Validation setting			Considering Validation setting		
Parameters	Estimate	SD	Parameters	Estimate	SD
θ_0	0.426	0.014	r_0	0.383	0.046
θ_1	0.537	0.018	r_1	0.605	0.052
			SN	0.700	0.043
			SP	0.733	0.041
$\log(OR)$	0.445	0.092	$\log(OR)$	0.912	0.233
95%C.I. (OR)	Does not include H_0 value (1.302, 1.870)		95%C.I. (OR)	Does not include H_0 value (1.654, 4.085)	

For the Bayesian estimates and the trace plots, the initial values of r_0, r_1, SN and SP were set to 0.4, 0.4, 0.7 and 0.7 respectively. For θ_0 and θ_1, it was 0.45 and 0.45.

The Gelman and Rubin convergence diagnostic statistic, $\hat{R}$ value for the four parameters r_0, r_1, SN and SP are 1.23, 1.16, 1.11 and 1.26 respectively. Also, for θ_0 and θ_0, $\hat{R}$ gives 1.26 and 1.28 respectively. Notice that, most of these values are over 1.2 for 10,000 iterations considering half of these as burn-in. Hence we can conclude that the convergence is not good for the cases under consideration for 10,000 iterations. If we increase the number of iterations to 40,000 and $\hat{R}$ value for the four parameters r_0, r_1, SN and SP becomes 1.19, 1.16, 1.05 and 1.15 respectively. For θ_0 and θ_0, $\hat{R}$ now gives 1.01 and 1.10 respectively. As all of these $\hat{R}$ values are less than 1.2, we can conclude that the convergence is satisfactory for the cases under consideration for 40,000 iterations. Therefore, we report the trace plots and the Bayesian estimates of the parameters for 40,000 iterations in Table 5.6 and Figure 5.3. However, it should be noted that the changes in estimation are very small (changes mostly in third decimal places) and the standard errors are almost the same despite the larger number of iterations.

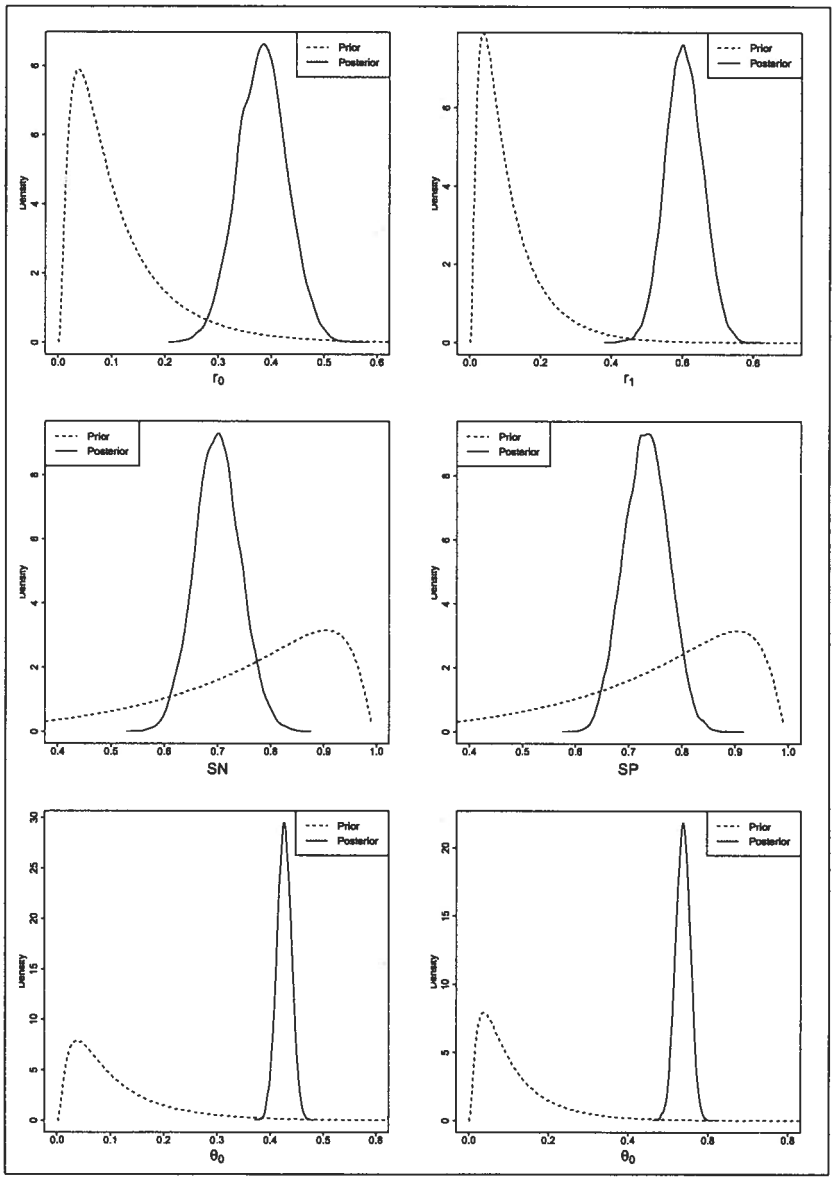
As before, the initial values were set to be 0.2, 0.4, 0.6 and 0.8 for each of the parameters under consideration. One possible reason for this analysis requiring such large number of iterations could be due to the fact that some

Figure 5.3: MCMC for the with and without validation data model parameters in the HSV-2 study



cell counts of the Table 5.4 are 5 or less. Another possibility is that the nondifferential assumption does not hold in this case.

Figure 5.4: Prior and Posterior Distributions of all the Parameters under Consideration in the HSV-2 study



Chapter 6

Conclusions and Further Research

6.1 Overall Conclusions

Various practical issues force researchers to use inferior measures of exposure assessment. When an ideal exposure measurement is replaced by an operational method or a surrogate variable, it is well known in the literature that due to the disparity between these two measures, there are several consequences of such compromise. Of course the extent of disparity plays a role in the consequences. To understand the extent to which the measure of association differs, a validation sub-sample is used to get some insight about the misclassification probabilities. Using the added information obtained from a validation sub-sample, adjustment measures are possible to correct for such bias and the subsequent power loss in hypothesis testing procedures.

The nondifferentiality assumption is very popular in the epidemiologic literature due to its various attractive features. Two adjustment techniques are considered in this thesis under this assumption. One is based on frequentist methods, power curves were derived for the likelihood ratio test both with and without validation data. This is basically a standard routine, used here as a benchmark. The detailed procedure is discussed in Chapter 2. The main goal is to evaluate the Bayesian counterpart which is based on a MCMC algorithm after reasonable diagnostic checks as discussed in Chapter 3. Both these methods are implemented in two settings: considering validation data and without considering validation data. In the frequentist method, estimates from the validation sub-sample are used to adjust for exposure misclassification, but in the Bayesian implementation, instead of having specific estimates of parameters, a set of priors are used so that some randomness or uncertainty is induced in the inferential process amongst cases and controls.

The main focus of this research is to identify the adjustment methodology that performs better under fairly general conditions in hypothesis testing. A set of scenarios are considered so that both methods can be compared using simulation study. These scenarios were constructed by varying the level of misclassification, prevalence, sample size, proportion of validation part in the whole sample and under fixed cost constraint. Since a lot of scenarios are in possible, to simplify the problem, only the one dimensional effects due to the one of the parameters, sample size or sample composition change is considered at a time. Both methods are applied on all of these scenarios. Details are provided in Chapter 4. As a tool of evaluation, power curves are drawn for the frequentist method and the proportion of credible intervals that exclude the null value are plotted for Bayesian method. From these plots, it is clear that the with validation data model is always better. The without validation data (two parameter) model can be as good as with validation data (four parameter) model in extreme cases, but can never get better. We showed that this is true for hypothesis testing settings. The only case when the without validation data model can be superior to the with validation data model is under fixed budget, if the cost of collecting validated data is much higher than collecting usual unvalidated data. How high is high? This depends on the various parameters, considered sample sizes, composition of sample and budget for the study. We just showed by example that such an exception is possible.

It is worth mentioning that the settings considered by *Greenland and Gustafson* [2006] are slightly different than those considered in this work, although they also address the issue of adjusting for misclassification in the context of hypothesis testing. In that paper, it is shown that given known or reasonably assumed (say, from educated guesses) values of sensitivity and specificity, the power does not improve after adjustment under nondifferential misclassification error (assumed to be free from any other sources of errors). This suggestion was based on the analysis of a single dataset. In contrast, in the current work, we showed that in presence of validation data, which enables us to estimate the true exposure prevalences, sensitivity and specificity, we have more power after adjustment, subject to the condition that the nondifferential misclassification assumption is satisfied.

If the plots of the frequentist and the Bayesian method results for respective scenario are superimposed, they are almost indistinguishable. Comparing these plots under each scenario, it is evident that both methods perform exactly the same way. Having both methods producing the same conclusion,

it is worth mentioning that the Bayesian framework, although very easy to generalize to other extensions of this problem, are very demanding in terms of resources and computing time to attain results without any MCMC diagnostic anomaly. On the other hand, with the frequentist methods used here, although closed forms are not always attainable, simple numerical routines can optimize these likelihoods very quickly. To give real life flavor, two epidemiologic datasets are also analyzed using the above methodologies in chapter 5, which are coherent with the simulation results.

6.2 Further Research and Recommendations

Further research could focus on extending some of the simplistic assumptions that were considered, adapting the proposed models for problems with similar specifications and generalizing the simulation scenario to broader contexts.

- One can consider larger combinations of the scenario setting than considered in this work to describe the effects in a broader sense. One could organize this effort by developing an experimental design (e.g., a fractional factorial design) involving the factors of interest.

- One immediate extension of the work is to go beyond nondifferential assumption and check the results under differential misclassification, which is more realistic in many fields. For Bayesian adjustment, this can be easily done by considering the general model where the misclassification probabilities are different with respect to case and control and imposing a joint prior for those parameters with an assumed covariance structure.

- To make the problem more realistic, additional exposures that are correctly measured are worth adding in the model. A logistic regression model can be a start in this direction.

- This dissertation only deals with dichotomous exposure misclassification. Polytomous exposure misclassification can also be another extension to this research. Instead of binomial assumption of misclassified exposure, a multinomial assumption will be used in that case.

- The models used in this work can be modified to allow using replicated sub-set of data or data obtained from an alternative source in the absence of a benchmark scorer or gold standard method of exposure assessment, instead of validation data, which could be more cost effective, especially when the cost of validation data is very high.

- It is also worth investigating other tools to analyze the continuous exposure data directly, instead of dichotomizing it to make it categorical, and try to identify how much sensitivity does one loose by categorizing the exposure variable. Spline analysis can be one way to move in this direction.

- The Bayesian hypothesis testing could be accomplished by using the Bayes factor, and then compared with the standard likelihood techniques to find out whether there is any discrepancy in those two methodologies.

Bibliography

Barron, B., The effects of misclassification on the estimation of relative risk, *Biometrics*, *33*(2), 414–418, 1977.

Brooks, S., P. Dellaportas, and G. Roberts, An approach to diagnosing total variation convergence of MCMC algorithms, *Journal of Computational and Graphical Statistics*, *6*(3), 251–265, 1997.

Bross, I., Misclassification in 2 x 2 tables, *Biometrics*, *10*(4), 478–486, 1954.

Broyden, C., The convergence of a class of double-rank minimization algorithms 1. general considerations, *IMA Journal of Applied Mathematics*, *6*(1), 76–90, 1970.

Carroll, R., M. Gail, and J. Lubin, Case-control studies with errors in covariates, *Journal of the American Statistical Association*, *88*(421), 185–199, 1993.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman & Hall/CRC, 2006.

Chen, T., A review of methods for misclassified categorical data in epidemiology, *Statistics in Medicine*, *8*(9), 914–921, 1989.

Chu, R., Bayesian Adjustment for Exposure Misclassification in Case-Control Studies, Master's thesis, Department of Statistics, University of British Columbia, 2005.

Copeland, K., H. Checkoway, A. McMichael, and R. Holbrook, Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology*, *105*(5), 488–495, 1977.

Cowles, M., and B. Carlin, Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association*, *91*(434), 883–904, 1996.

Diamond, E., and A. Lilienfeld, Effects of errors in classification and diagnosis in various types of epidemiological studies, *American Journal of Public Health*, *52*(7), 1137–1145, 1962a.

Diamond, E., and A. Lilienfeld, Misclassification errors in 2 x 2 tables with one margin fixed: some further comments, *American Journal of Public Health*, *52*(12), 2106–2111, 1962b.

Dosemeci, M., S. Wacholder, and J. Lubin, Does nondifferential misclassification of exposure always bias a true effect toward the null value?, *American Journal of Epidemiology*, *132*(4), 746–748, 1990.

Fletcher, R., A new approach to variable metric algorithms, *The Computer Journal*, *13*(3), 317–322, 1970.

Gelman, A., *Bayesian Data Analysis*, CRC Press, 2004.

Gelman, A., and D. Rubin, Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*(4), 457–472, 1992.

Geyer, C., Practical markov chain monte carlo, *Statistical Science*, *7*(4), 473–483, 1992.

Gill, J., *Bayesian Methods: A Social and Behavioral Sciences Approach*, Chapman & Hall/CRC, 2008.

Gladen, B., and W. Rogan, Misclassification and the design of environmental studies, *American Journal of Epidemiology*, *109*(5), 607–616, 1979.

Goldberg, J., The Effects of Misclassification on the Analysis of Data in 2 X 2 Tables, *Unpublished dissertation, Harvard University School of Public Health. Boston*, 1972.

Goldberg, J., The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table, *Journal of the American Statistical Association*, *70*(351), 561–567, 1975.

Goldfarb, J., A family of variable metric updates derived by variational means, *Mathematics of Computing*, *24*(109), 23–26, 1970.

Greenland, S., The effect of misclassification in the presence of covariates, *American Journal of Epidemiology*, *112*(4), 564–569, 1980.

Greenland, S., Statistical uncertainty due to misclassification: implications for validation substudies., *Journal of Clinical Epidemiology*, *41*(12), 1167–1174, 1988a.

Greenland, S., Variance estimation for epidemiologic effect estimates under misclassification, *Statistics in Medicine*, *7*(7), 745–757, 1988b.

Greenland, S., Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification, *Journal of Statistical Planning and Inference*, *138*(2), 528–538, 2008.

Greenland, S., and P. Gustafson, Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction, *American Journal of Epidemiology*, *164*(1), 63–68, 2006.

Gullen, W., J. Bearman, and E. Johnson, Effects of misclassification in epidemiologic studies., *Public Health Reports*, *83*(11), 914–918, 1968.

Gustafson, P., *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, CRC Press, 2004.

Gustafson, P., and S. Greenland, Curious phenomena in Bayesian adjustment for exposure misclassification, *Statistics in Medicine*, *25*(1), 2006.

Gustafson, P., N. Le, and R. Saskin, Case-control analysis with partial knowledge of exposure misclassification probabilities, *Biometrics*, *57*(2), 598–609, 2001.

Hastings, W., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, *57*(1), 97–109, 1970.

Joseph, L., T. Gyorkos, and L. Coupal, Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology*, *141*(3), 263–272, 1995.

Jurek, A., S. Greenland, G. Maldonado, and T. Church, Proper interpretation of non-differential misclassification effects: expectations vs observations, *International journal of epidemiology*, *34*(3), 680–687, 2005.

Keys, A., and J. Kihlberg, Effect of misclassification on estimated relative prevalence of a characteristic: Part I. Two populations infallibly distinguished. Part II. Errors in two variables, *American Journal of Public Health*, *53*(10), 1656, 1963.

Koch, G., The Effect of Non-Sampling Errors on Measures of Association in 2× 2 Contingency Tables, *Journal of the American Statistical Association*, *64*(327), 852–863, 1969.

Kraus, J., S. Greenland, and M. Bulterys, Risk factors for sudden infant death syndrome in the US Collaborative Perinatal Project, *International Journal of Epidemiology*, *18*(1), 113–120, 1989.

Lyles, R., A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure, *Biometrics*, *58*(4), 1034–1037, 2002.

Marshall, J., The use of dual or multiple reports in epidemiologic studies, *Statistics in Medicine*, *8*(9), 1041–1049, 1989.

Marshall, R., Validation study methods for estimating exposure proportions and odds ratios with misclassified data., *Journal of Clinical Epidemiology*, *43*(9), 941–947, 1990.

Marshall, R., Assessment of exposure misclassification bias in case-control studies using validation data, *Journal of Clinical Epidemiology*, *50*(1), 15–19, 1997.

Morrissey, M., and D. Spiegelman, Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons, *Biometrics*, *55*(2), 338–344, 1999.

Newell, D., Errors in the interpretation of errors in epidemiology, *American Journal of Public Health*, *52*, 1925–1928, 1962.

Prescott, G., and P. Garthwaite, A simple Bayesian analysis of misclassified binary data with a validation substudy, *Biometrics*, *58*(2), 454–458, 2002.

Raftery, A., and S. Lewis, How many iterations in the Gibbs sampler, *Bayesian Statistics*, *4*, 763–773, 1992.

Rahme, E., L. Joseph, and T. Gyorkos, Bayesian sample size determination for estimating binomial parameters from data subject to misclassification, *Applied Statistics*, *49*(1), 119–128, 2000.

Rizzo, M., *Statistical Computing with R*, Taylor & Francis, Inc., 2007.

Robert, C., Convergence control methods for Markov chain Monte Carlo algorithms, *Statistical Science*, *10*(3), 231–253, 1995.

Rothman, K., S. Greenland, and T. Lash, *Modern Epidemiology*, Lippincott Williams & Wilkins, 2008.

Shanno, D., Conditioning of quasi-Newton methods for function minimization, *Mathematics of Computation*, *24*(111), 647–656, 1970.

Thomas, D., When will nondifferential misclassification of an exposure preserve the direction of a trend?, *American Journal of Epidemiology*, *142*(7), 782–784, 1995.

Walter, S., and L. Irwig, Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review., *Journal of Clinical Epidemiology*, *41*(9), 923–937, 1988.

Willett, W., An overview of issues related to the correction of nondifferential exposure measurement error in epidemiologic studies, *Statistics in Medicine*, *8*(9), 1031–1040, 1989.

Youden, W., Index for rating diagnostic tests, *Cancer*, *3*(1), 32–35, 1950.