

**Case-control Studies with Misclassified
Exposure: A Bayesian Approach**

by

Refik Saskin

B.Sc, Brock University, 1998

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

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

August 2000

© Refik Saskin, 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date AUG 09, 2000.

Abstract

When dealing with the case-control data, it is often the case that the exposure to a risk factor of interest is subject to misclassification. Methods for correcting the odds-ratio are available when the misclassification probabilities are known. In practice, however, good guesses rather than the exact values are available for these probabilities. We show that when these guesses are treated as exact even the smallest differences between the true and guessed values can lead to very erroneous odds-ratio estimates. This problem is alleviated by a Bayesian analysis which incorporates the uncertainty about the misclassification probabilities as prior information.

In practice, data on the exposure variable are quite often available from more than one source. We review three methods for improving the odds-ratio estimates that combine information from two sources. We then develop a Bayesian approach which is based on latent class analysis, and apply it to the sudden infant death syndrome data.

The inference required the use of the Metropolis-Hastings algorithm and/or the Gibbs sampler.

Contents

Abstract	ii
Contents	iii
List of Tables	v
List of Figures	vii
Acknowledgements	xii
Dedication	xiii
1 Introduction	1
2 One Test Case	5
2.1 Definitions and Terminology	5
2.1.1 Data and Setup	7
2.2 Known Sensitivity and Specificity	10
2.3 The Gibbs Sampler	17
2.3.1 Examples	21

2.4	The Metropolis-Hastings Algorithm	24
2.4.1	Large Sample Case	31
2.4.2	Finite Sample Case	35
2.4.3	Examples	36
2.5	Discussion	40
3	Two Test Case	49
3.1	Data and Setup	49
3.2	Correction Methods	51
3.2.1	The Marshall and Graham Method	51
3.2.2	The Drews, Flanders and Kosinski Method	52
3.2.3	The Kaldor and Clayton Method	54
3.3	Our Method	56
3.3.1	Examples	60
3.4	An Application to Real Data	68
3.5	Discussion	72
4	Conclusion	74
	Bibliography	78

List of Tables

2.1	Distribution of subjects in a case-control study by disease status and an imperfect measurement of a dichotomous exposure. Symbol + denotes apparently exposed, and – denotes apparently unexposed.	9
2.2	(a) True distribution of exposure among cases and controls. (b) Observed distribution of exposure, given 10% misclassification among cases and controls. (c) Estimated distribution of exposure assuming 14% misclassification among cases and controls.	14
2.3	Distribution of observed and latent data when one imperfect classification procedure is used. Symbol + denotes exposed, and – denotes unexposed.	18
3.1	Distribution of subjects in a case-control study when two tests are used to assess the exposure status. Symbol + denotes exposed and – denotes unexposed.	50

3.2	Distribution of observed and latent data when two imperfect tests are used, together with the contribution to the likelihood each combination of the observed and latent data makes. Symbol $+$ denotes exposed, and $-$ denotes unexposed.	57
3.3	True parameter values and the empirical coverage of the 80% HPD credible intervals for $\log \phi$ for the sample sizes $N_i = 200$, $N_i = 800$, and $N_i = 3200$, $i = 0, 1$. Numbers in brackets represent mean interval length. For the empirical coverage and the mean length, 1000 data sets were simulated.	67
3.4	A pseudorandom sample of 226 SIDS cases and 226 controls from the NICHD study. Data was classified using medical record (MR) and interview (Int) data.	69
3.5	Median estimates of the sensitivities, specificities and the logarithm of the odds-ratio, and 95% HPD credible intervals for the logarithm of the odds-ratio, using the Gibbs sampler.	73
3.6	95% HPD credible intervals for the sensitivities and specificities of the interview data and the medical records.	73
3.7	Estimates of the sensitivities, specificities and the logarithm of the odds-ratio, and 95% confidence intervals for the logarithm of the odds-ratio, using the EM algorithm (Drews et al. (1993)).	73

List of Figures

2.1	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 200$	25
2.2	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 800$	26
2.3	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 3200$	27
2.4	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 200$	28

2.5	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 800$	29
2.6	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 3200$	30
2.7	The support of $p(\alpha, \beta \theta_0, \theta_1)$. The shaded rectangles comprise $A(\theta_0, \theta_1)$	32
2.8	Posterior distribution of $\log \phi$ in Example 1. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$	42
2.9	80% highest posterior density credible intervals for $\log \phi$ in Example 1. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.	43

2.10	Prior and posterior samples of the sensitivity and specificity for the datasets in Example 1.	44
2.11	Posterior distribution of $\log \phi$ in Example 2. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$	45
2.12	80% highest posterior density credible intervals for $\log \phi$ in Example 2. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.	46
2.13	Posterior distribution of $\log \phi$ in Example 3. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$	47

2.14	80% highest posterior density credible intervals for $\log \phi$ in Example 3. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.	48
3.1	Post burn-in output of the five independent chains of the Gibbs sampler for sensitivities, specificities, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 200$	62
3.2	Posterior samples of the logarithm of the odds-ratio for the datasets in Example 1.	64
3.3	80% highest posterior density credible intervals for $\log \phi$ in Example 1. The solid vertical lines represent credible intervals for forty data sets and the solid horizontal line indicates the true value of $\log \phi$. The panel (a) gives the intervals for the sample size $N_0 = N_1 = 200$, the panel (b) gives the intervals for the sample size $N_0 = N_1 = 800$, and the panel (c) gives the intervals for the sample size $N_0 = N_1 = 3200$	65

3.4	Posterior samples of the logarithm of the odds-ratio for the sample size $N_0 = N_1 = 800$ in Example 1. (a) $Beta(1, 1)$ prior used for all six parameters. (b) $Beta(7.55, 2.64)$ prior used for α_i and $\beta_i, i = 1, 2$, while $Beta(1, 1)$ prior used for π_0 and π_1 (the histogram is identical to that shown in Figure 3.2 (b)). (c) $Beta$ prior for each parameter was chosen so that the mode γ is centred at the true value and 0.95 probability is assigned to the interval $\gamma \pm 0.05$	66
-----	--	----

Acknowledgements

I would like to thank Dr. Paul Gustafson for his consistent guidance, help and support. His patience and expert advice made the completion of my research possible. I would also like to thank Dr. Nhu Lee for his generous input and willingness to share his expertise with me.

Many thanks to the entire UBC Statistics Department for making these last two years most pleasurable.

Also, many thanks to my family for all their love, support and encouragement. Finally, I would like to thank my girlfriend Marnie for her continued support and patience.

REFIK SASKIN

The University of British Columbia

August 2000

To my mother, God rest her soul.

Chapter 1

Introduction

Misclassification of exposure is one of the most serious problems in epidemiology. Even on the smallest scale, exposure misclassification can substantially bias estimates of the relative risk. In particular, nondifferential misclassification of a dichotomous exposure variable always tends to attenuate observed exposure-disease relationships. This attenuation can be particularly large for rare exposures and imperfect specificity or highly prevalent exposure and imperfect sensitivity of the test used to assess the exposure status. The corrected estimates can be obtained easily if the sensitivity and specificity of the classification procedure are known, which was illustrated by Barron (1977) and Greenland and Kleinbaum (1983). In most cases, however, these quantities are very difficult to estimate, because no measure of true exposure is available, and even if available, such "gold standard" measures are often very expensive and/or invasive. A more common scenario is when good guesses rather than exact values are available for the sensitivity and specificity. If one treats these

guesses as exact, Marshall (1989) has shown that even small discrepancies between the true and guessed values of the sensitivity and specificity can lead to very erroneous odds-ratio estimates. The effects of ignoring misclassification and the methods which correct for it have received a considerable attention in the literature. Thomas, Stam and Dwyer (1993) and Walter and Irwig (1988) provide reviews.

Although one can rarely obtain completely accurate exposure data for epidemiologic studies, data on the exposure variable are often available from more than one source. For instance, exposure data are often available from both medical records and interviews, or from two different diagnostic tests performed on a study subject. A variety of methods have been suggested for improving the odds-ratio estimates by combining data from two information sources. One simple approach, as suggested by Marshall and Graham (1984), is to restrict the analysis to subjects for whom the two sources agree about the exposure. As pointed out by Walter (1984), however, this method may result in a substantial loss of precision due to the exclusion of a potentially large proportion of study subjects. Hui and Walter (1980) developed a method for estimating the sensitivity and specificity of two classification schemes in two groups (such as controls and cases) when, given the true exposure status, misclassification is nondifferential and the two schemes classify study subjects independently. More recent work by Drews, Flanders and Kosinski (1993) has extended their method to more complicated settings. They perform numeric maximization rather than providing a closed-form solution. The approach of

Drews et al. (1993) requires that the analyst specify the degree of dependency between the two classification schemes. Since, in practice, a gold standard may be unavailable, impractical, or itself measured with error, this method could have limited applicability. In their discussion of latent class analysis, Kaldor and Clayton (1985) give an example of data where replicate measurements are available for some or all of the cases and/or controls. They demonstrate that obtaining a replicate measurements on even a modest proportion of subjects leads to substantially improved estimation of odds-ratio.

In the following chapter we will consider a case-control setting where one imperfect classification scheme is used to assess the exposure status. In this context, we will introduce a simple approach to Bayesian inference about the odds-ratio when the misclassification probabilities are known. Then, we will outline a Bayesian approach to odds-ratio inference with only partial knowledge of classification probabilities. Two computational methods will be presented that enable the inference on odds-ratio. Finally, examples will be presented illustrating the use of these methods.

In chapter 3 we will look at a case-control setting where the exposure status is measured by two imperfect classification schemes. We will briefly review three existing methods to correct the odds-ratio estimates. Then, we will introduce a Bayesian approach. It is an extension of the Joseph, Gyorkos and Coupal (1995) method for inference on the sensitivities and specificities of two classification procedures applied to one population. We will examine the validity of our method by applying it to nine hypothetical case-control studies,

where the true odds-ratio are known. Finally, we will apply our method to a case-control study of sudden infant death syndrome, and compare the results with those obtained using the method of Drews et al. (1993).

Chapter 2

One Test Case

2.1 Definitions and Terminology

In this section we present some key definitions and terminology, making the subsequent sections easier to follow. In a “case-control study” investigator selects “cases” of a disease, and a comparison group, called “controls”. The cases and controls are then compared with respect to a specific risk factor, often referred to as “exposure”. The samples of the cases and controls are usually regarded as independent.

“Misclassification” or “measurement error” refers to any difference between the true value of a variable and its measured value. The term misclassification is almost always used in the context of categorical variables and measurement error when one refers to continuous variables. Errors can be “random” or “systematic”. If the error is not randomly distributed around its true value, we say that it is systematic. Overestimating everyone’s exposure

level by a factor of two would be an example of a systematic error.

Systematic and random errors can be either "differential" or "nondifferential". If the misclassification in the exposure depends upon the disease status, then the misclassification is differential. If on the other hand, the misclassification in the exposure does not depend on the disease status, we say that the misclassification is nondifferential. A more rigorous definition of nondifferential misclassification is that the risk of disease depends only on the true exposure, and given this true exposure, the measured exposure does not add any additional information.

In clinical medicine and epidemiology, tests are often used to determine the presence or absence of a disease, or whether or not someone was exposed to a risk factor of interest. Ideally, those who are exposed (or with disease) should be classified as being exposed (or with disease), and those who are not exposed (or without disease) should be classified as unexposed (or without disease). The "sensitivity" and "specificity" of a test consider how often such correct classification occurs. The sensitivity of a test is the percentage of exposed study subjects (or with disease) classified as exposed (or with disease). The specificity of a test is the percentage of unexposed study subjects (or without disease) classified as unexposed (or without disease). Moreover, the "false positive rate" (FPR) of a test is the percentage of unexposed study subjects (or without disease) who are classified as exposed (or with disease). The "false negative rate" (FNR) of a test is the percentage of exposed study subjects (or with disease) who are classified as unexposed (or without disease). Therefore,

$\text{FPR} = 1 - \text{specificity}$ and $\text{FNR} = 1 - \text{sensitivity}$.

As the methodology we develop in subsequent chapters is of Bayesian nature, we will very briefly introduce some of its key features. In a nutshell, the Bayesian approach uses probability to describe both model parameters and random variables. In the context of the usual statistical model with a random variable X having possible distributions indexed by parameter θ , the data x becomes known to the statistician and the object is to make inference about the unknown parameter. The information about θ that is available to the statistician prior to observing the data and his/her belief about the parameter are reflected in a “prior distribution”. In the Bayesian approach, the statistician will wish to calculate the probability distribution of θ given $X = x$, called the “posterior distribution”. Once this is accomplished, point and interval estimates of θ can be calculated, significance tests can be performed, etc. With the recent development of computer techniques has come an increase in the popularity of Bayesian inference.

2.1.1 Data and Setup

Suppose we have a case-control study attempting to assess the relationship between disease status and a dichotomous exposure (usually labeled as 0 for unexposed and 1 for exposed). Suppose further that the measurement procedure used to assess the exposure status is not perfect, i.e., the sensitivity and specificity of that procedure are less than one (so that there is a non-negligible number of false positives and false negatives). The study, therefore, consists of

measuring the apparent exposure status for random sample of N_0 controls and N_1 cases. We will assume that the random samples are independent and that the disease status is known exactly, without error. These assumptions are not particularly unreasonable or limiting. First, the independence between cases and controls is a usual assumption one makes in a case-control study. Second, in order to determine the potential treatment, clinicians concentrate on correctly identifying the disease. For a disease that is difficult to diagnose, multiple tests, sometimes even invasive procedures, are applied. Moreover, the diseased subjects chosen to participate in a case-control study have usually had the condition for quite some time before the start of a study, increasing the chance of correct diagnosis. As a result, the disease status is (usually) very well determined.

Throughout this thesis we will assume that the misclassification is non-differential, i.e., independent of disease status.

In the case-control setting, the data comes in a two-by-two table, as shown in Table 2.1. Here, x_i and y_i are the observed number of positive and negative test results, respectively, in the sample of $x_i + y_i = N_i$ study subjects. The subscript $i = 0$ denotes controls, $i = 1$ cases.

Furthermore, let D , E and E' represent disease status, actual exposure status and apparent exposure status, respectively. In each case, suppose that one represents presence, and zero represents absence. As nondifferential misclassification is assumed, the relevant parameters are

$$\pi_0 = Pr(E = 1 | D = 0),$$

	Test		
	+	-	
Controls	x_0	y_0	N_0
Cases	x_1	y_1	N_1
			N

Table 2.1: Distribution of subjects in a case-control study by disease status and an imperfect measurement of a dichotomous exposure. Symbol + denotes apparently exposed, and - denotes apparently unexposed.

$$\pi_1 = Pr(E = 1|D = 1),$$

$$\alpha = Pr(E' = 1|E = 1),$$

$$\beta = Pr(E' = 0|E = 0).$$

Here, α is the sensitivity of the classification procedure, β is the specificity of the classification procedure, and π_0 and π_1 are the exposure prevalences among the controls and cases, respectively. The false positive and false negative rates are $1 - \beta$ and $1 - \alpha$, respectively. The odds-ratio, denoted as ϕ , is defined in terms of the exposure prevalences as

$$\phi = \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)}.$$

In general, suppose that R tests are applied to S populations and that misclassification is nondifferential. As R tests are used to determine the exposure status, there are 2^R possible classification outcomes. Furthermore, since

the number of subjects is fixed for each population, there are $2^R - 1$ degrees of freedom per population, giving a total of $(2^R - 1)S$ degrees of freedom available. On the other hand, if the misclassification is nondifferential, there are $2R + S$ parameters to estimate, the sensitivity and specificity of each test and S exposure prevalences. Therefore, in the case of one test and two population ($R = 1, S = 2$), the likelihood with no constraint placed on the parameters is overparameterized. Hence, if $(\alpha, \beta, \pi_0, \pi_1)$ are unknown, the resulting likelihood function from a case-control study where one classification scheme is used to assess the exposure is nonidentifiable (Hui and Walter (1980)).

2.2 Known Sensitivity and Specificity

If the misclassification probabilities α and β are known, Barron (1977) and Greenland and Kleinbaum (1983) present a relatively straightforward procedure for the adjustment of odds-ratio estimates. To illustrate their methodology, suppose that α and β are the sensitivity and specificity of a test used to assess the exposure, while α^* and β^* are the sensitivity and specificity of a test used to assess the disease in a case-control study. Note that their method does not require the assumption that the disease status is known exactly. Suppose further that (x_{Ti}, y_{Ti}) and (x_{Mi}, y_{Mi}) are, respectively, true and misclassified cell counts in a 2×2 table (see Table 2.1). When there is no misclassification, we estimate the odds-ratio by

$$\widehat{\phi}_T = \frac{x_{T1}y_{T0}}{x_{T0}y_{T1}}, \quad (2.1)$$

and if the exposure is subject to misclassification, we estimate the odds-ratio by

$$\widehat{\phi}_M = \frac{x_{M1}y_{M0}}{x_{M0}y_{M1}}. \quad (2.2)$$

The misclassified cell counts are related to the true cell counts by the set of four equations $\mathbf{m} = \mathbf{C}\mathbf{t}$, where

$$\mathbf{m} = \begin{bmatrix} x_{M0} \\ y_{M0} \\ x_{M1} \\ y_{M1} \end{bmatrix}, \mathbf{t} = \begin{bmatrix} x_{T0} \\ y_{T0} \\ x_{T1} \\ y_{T1} \end{bmatrix},$$

and

$$\mathbf{C} = \begin{bmatrix} \alpha\beta^* & (1-\beta)\beta^* & \alpha(1-\alpha^*) & (1-\beta)(1-\alpha^*) \\ (1-\alpha)\beta^* & \beta\beta^* & (1-\alpha)(1-\alpha^*) & \beta(1-\alpha^*) \\ \alpha(1-\beta^*) & (1-\beta)(1-\beta^*) & \alpha\alpha^* & (1-\beta)\alpha^* \\ (1-\alpha)(1-\beta^*) & \beta(1-\beta^*) & (1-\alpha)\alpha^* & \beta\alpha^* \end{bmatrix}.$$

Therefore, if $(\alpha, \beta, \alpha^*, \beta^*)$ are known and \mathbf{C} is invertible, a correction formula is easily derived, since $\mathbf{t} = \mathbf{C}^{-1}\mathbf{m}$. This, in turn, enables us to compute the estimate of the true odds-ratio ϕ_T .

We now present a simple approach to Bayesian inference about odds-ratio when the sensitivity and specificity are known.

Let X_0 be the number of apparently exposed controls and X_1 be the number of apparently exposed cases. A control can be truly exposed and correctly classified as exposed with probability $(\pi_0\alpha)$ or unexposed and incorrectly classified as exposed with probability $(1-\pi_0)(1-\beta)$. Similar is true for the

cases. Therefore, X_0 and X_1 are distributed as independent $Binomial(N_i, \theta_i)$ random variables, where

$$\theta_i = \pi_i \alpha + (1 - \pi_i)(1 - \beta), i = 0, 1. \quad (2.3)$$

Here, θ_0 and θ_1 are apparent exposure prevalences among controls and cases, respectively.

Suppose now that α and β are known exactly, without error. Further, note that an estimate of ϕ can be obtained from estimates of θ_0 and θ_1 , as

$$\begin{aligned} \phi &= \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)} \\ &= \frac{(\theta_1 + \beta - 1)/(\alpha - \theta_1)}{(\theta_0 + \beta - 1)/(\alpha - \theta_0)}. \end{aligned} \quad (2.4)$$

Thus, inference about the odds-ratio ϕ is simple. To illustrate this, suppose that independent $Uniform[0, 1]$ priors are chosen for π_0 and π_1 , i.e., no prior knowledge is assumed about the prevalence of exposure in controls and cases. This choice of prior implies $Uniform[\min(1 - \beta, \alpha), \max(1 - \beta, \alpha)]$ priors for θ_0 and θ_1 , since (2.3) implies $\theta_i \in [\min(1 - \beta, \alpha), \max(1 - \beta, \alpha)]$, $i = 0, 1$. Consequently, $\theta_i | x_0, x_1$ follows a $Beta(x_i + 1, N_i - x_i + 1)$ distribution truncated to the interval $[\min(1 - \beta, \alpha), \max(1 - \beta, \alpha)]$, yielding the posterior distribution of $\phi | x_0, x_1$. Therefore, sampling from the posterior distribution of odds-ratio is straightforward. Examples illustrating this approach will be shown in section 2.4.3.

We now examine a more realistic scenario, when good guesses rather than exact values are available for the sensitivity and specificity. These guesses could be available, for instance, via previous studies. If these guesses are treated as exact, Marshall (1989) has demonstrated that even small differences between the guessed and true values of misclassification probabilities can lead to very erroneous odds-ratio estimates. More specifically, consider the following example. Table 2(a) below shows the true distribution of exposure among 300 cases and 300 controls. In this case, $\log \phi = 1.35$, with a 95% confidence interval of (0.90, 1.80). This confidence interval was computed assuming the asymptotic normality of $\log \phi$. Table 2(b) presents the effect the misclassification of 10% has on the odds-ratio. Here, $\log \phi = 0.85$, with a 95% confidence interval of (0.47, 1.23), thus lessening the apparent effect of exposure. Of the 30 exposed controls, only 90% are correctly identified as exposed and 10% are mistakenly identified as not exposed.

Ignoring the sampling error, of the 270 unexposed controls, 90% are correctly classified with 10% classified as exposed. Hence, the 54 controls classified as exposed in Table 2(b) include 90% of 30, or 27, who are exposed and 10% of 270, or 27, who are unexposed. The same process applies to cases.

The investigator who knows that the misclassification is 10% can easily adjust the data in Table 2(b) and correctly estimate the odds-ratio. A slight miscorrection can yield an erroneous estimate of the odds-ratio. To illustrate this, suppose that the investigator may guess, on the basis of previous studies, that the misclassification is 14%. Table 2(c) displays the result of adjusting

(a)		Test		
		+	-	
Controls	30	270	300	
Cases	90	210	300	
				600

(b)		Test		
		+	-	
Controls	54	246	300	
Cases	102	198	300	
				600

(c)		Test		
		+	-	
Controls	17	283	300	
Cases	83	217	300	
				600

Table 2.2: (a) True distribution of exposure among cases and controls. (b) Observed distribution of exposure, given 10% misclassification among cases and controls. (c) Estimated distribution of exposure assuming 14% misclassification among cases and controls.

the data in Table 2(b) using the guessed value. Here, $\log \phi = 1.85$, with a 95% confidence interval of (1.30, 2.40), thus producing an exaggerated estimate of the effect of exposure.

As shown in the above example, odds-ratio estimates can be very sensitive to small discrepancies between the actual and assumed values of α and β . We investigate this further by looking at the asymptotic bias of the odds-ratio that arises when incorrect values of the sensitivity and specificity are assumed.

To that end, let α' and β' be the assumed values of α and β used to correct the odds-ratio estimate. We say that “miscorrection” occurs when $(\alpha', \beta') \neq (\alpha, \beta)$. Moreover, we say that miscorrection is “asymptotically undetectable” if both θ_0 and θ_1 lie in the interval $[\min(1 - \beta', \alpha'), \max(1 - \beta', \alpha')]$. On the other hand, we say that miscorrection is “asymptotically detectable” if α' and β' are such that one or both of θ_0 and θ_1 lie outside the interval $[\min(1 - \beta', \alpha'), \max(1 - \beta', \alpha')]$. This distinction comes from the fact that, in the large sample case, the values of θ_0 and θ_1 are effectively known exactly. So, no amount of data could detect that miscorrection occurs if $\theta_i \in [\min(1 - \beta', \alpha'), \max(1 - \beta', \alpha')]$, $i = 0, 1$. Therefore, as N_0 and N_1 increase, the exposure prevalences converge to $\pi'_i = (\theta_i + \beta' - 1)/(\alpha + \beta - 1)$. Consequently, the posterior distribution of ϕ concentrates at

$$\phi' = \frac{(\theta_1 + \beta' - 1)/(\alpha' - \theta_1)}{(\theta_0 + \beta' - 1)/(\alpha' - \theta_0)}. \quad (2.5)$$

The difference $|\phi' - \phi|$ is the asymptotic bias. The following theorem illustrates potentially very dangerous consequences of a very small miscorrection.

Theorem: Suppose α and β are fixed, with $\alpha + \beta > 1$. Let $\epsilon_\alpha > 0$, $\epsilon_\beta > 0$ and $R > 0$ be arbitrary. Then, there exist α', β', π_0 and π_1 such that

- (i) there is asymptotically undetectable miscorrection,
- (ii) $|\alpha' - \alpha| < \epsilon_\alpha$, $|\beta' - \beta| < \epsilon_\beta$ and
- (iii) $\phi'/\phi = R$.

Proof: Note first that $\phi' = R\phi$, where

$$R = \frac{[1 + c/\pi_1] / [1 + d/(1 - \pi_1)]}{[1 + c/\pi_0] / [1 + d/(1 - \pi_0)]}, \quad (2.6)$$

$c = (\beta' - \beta)/(\alpha + \beta + 1)$ and $d = (\alpha' - \alpha)/(\alpha + \beta + 1)$. Note also that the requirement for asymptotic undetectability is equivalent to $-c \leq \pi_i \leq 1 + d, i = 0, 1$. Now, choose α' so that $|\alpha' - \alpha| < \epsilon_\alpha$ and $d > 1$ and choose β' such that $|\beta' - \beta| < \epsilon_\beta$ and $c \in (-1, 0)$. Therefore, asymptotic undetectability holds for $|c| \leq \pi_i \leq 1, i = 0, 1$. So now equation (2.6) is given by

$$R = \frac{[1 - |c|/\pi_1] / [1 + d/(1 - \pi_1)]}{[1 - |c|/\pi_0] / [1 + d/(1 - \pi_0)]}. \quad (2.7)$$

Hence, R increases to infinity as π_0 decreases to $|c|$ (or π_1 increases to one) and decreases to zero as π_1 decreases to $|c|$ (or π_0 increases to one), as required. Here we considered the case where $\alpha' > \alpha$ and $\beta' > \beta$. It is easy to establish the claim for other three cases.

It is worth noting that R can go to either zero or infinity without π_0 or π_1 going to either zero or one, which are rather unrealistic cases. Secondly, the assumption that $\alpha + \beta > 1$ is not unreasonable, given that no case-control study would be carried out if the false positive and false negative rates of the classification procedure are very high (higher than 0.5, say).

The previous work, coupled with the result of the above theorem, suggests that it seems reasonable to include the uncertainty in the available guesses into the analysis. Bayesian methods make this feasible, as the lack of exact knowledge of α and β is easily incorporated via the appropriate choice of priors for α and β . This methodology was introduced by Joseph, Gyorkos and Coupal (1995) in the context of applying R different diagnostic tests to S populations. Specifically, they focused their attention on one population and

one test ($R = 1, S = 1$) and one population and two tests ($R = 2, S = 1$). Consequently, their inference was on the sensitivity, specificity and disease prevalence of the population of interest. We will attempt to extend their approach to the case-control setting. The Gibbs sampler will be used to draw the posterior samples of α , β , π_0 and π_1 and consequently of ϕ .

2.3 The Gibbs Sampler

Let A_i and B_i be the information that is missing when the test used to assess the exposure is imperfect, that is the number of true positive test results out of a_i and b_i , respectively. Thus, A_i is the number of true positives and B_i is the number of false negatives, $i = 0, 1$. This missing information is called “latent data” and analysis of such data, called “latent class analysis”, has been done by Kaldor and Clayton (1985) and Walter and Irwig (1988). Incorporating this latent information into Table 2.1, we have (Table 2.3).

Using the independence between cases and controls, the likelihood function of the data in Table 2.3 is given by

$$\begin{aligned}
 l(A_0, B_0, A_1, B_1, x_0, y_0, x_1, y_1 | \pi_0, \pi_1, \alpha, \beta) &= \prod_{i=0,1} \frac{N_i}{A_i! \dots (y_i - B_i)!} \\
 &\times [\pi_i(1 - \alpha)]^{B_i} [(1 - \pi_i)\beta]^{y_i - B_i} \\
 &\times [\pi_i\alpha]^{A_i} [(1 - \pi_i)(1 - \beta)]^{x_i - A_i}.
 \end{aligned}$$

		Controls					Cases		
		True Exposure					True Exposure		
		+	-				+	-	
Apparent exposure	+	A_0	$x_0 - A_0$	x_0	+		A_1	$x_1 - A_1$	x_1
	-	B_0	$y_0 - B_0$	y_0	-		B_1	$y_1 - B_1$	y_1
				N_0					N_1

Table 2.3: Distribution of observed and latent data when one imperfect classification procedure is used. Symbol + denotes exposed, and - denotes unexposed.

By gathering the like terms, the likelihood becomes

$$\begin{aligned}
l(A_0, B_0, A_1, B_1, x_0, y_0, x_1, y_1 | \pi_0, \pi_1, \alpha, \beta) &= \prod_{i=0,1} \left[\frac{N_i}{A_i! \dots (y_i - B_i)!} \right. \\
&\times \pi_i^{A_i+B_i} (1 - \pi_i)^{N_i-A_i-B_i} \Big] \\
&\times \alpha^{A_0+A_1} (1 - \alpha)^{B_0+B_1} \\
&\times \beta^{y_0+y_1-B_0-B_1} (1 - \beta)^{x_0+x_1-A_0-A_1}
\end{aligned} \tag{2.8}$$

The likelihood (2.8) is nonidentifiable if $(\alpha, \beta, \pi_0, \pi_1)$ are unknown. We now state a more formal definition of nonidentifiability, introduced by Dawid (1979).

Suppose that the Bayesian model is denoted by likelihood $l(x|\Lambda)$ and prior $p(\Lambda)$, where $\Lambda = (\lambda_1, \lambda_2)$. We say that λ_2 is nonidentifiable if $p(\lambda_2|\lambda_1, x) = p(\lambda_2|\lambda_1)$. In other words, λ_2 is not identified by the data if observing data x

does not increase our knowledge about λ_2 given λ_1 . However, nonidentifiability does not imply that there is no Bayesian updating, i.e., it does not imply that $p(\lambda_2|x) = p(\lambda_2)$. Furthermore, since $p(\lambda_2|\lambda_1, x) \propto l(x|\Lambda)p(\lambda_2|\lambda_1)p(\lambda_1)$, λ_2 is nonidentifiable if and only if $l(x|\Lambda)$ is free of λ_2 . Hence, the definition of nonidentifiability, as introduced by Dawid (1979), is equivalent to nonidentifiability in the likelihood.

We now turn our attention to the choice of prior for α , β , π_0 and π_1 . We will choose beta densities to represent the prior information available for α , β , π_0 and π_1 . The reason for this choice of prior is three-fold. First, the beta density is positive on the interval $[0, 1]$, which coincides with the range of all parameters of interest. Secondly, the family of beta densities is flexible, in the sense that a variety of shapes can be chosen by selecting different values of the hyperparameters. Finally, it is the conjugate prior distribution for the binomial likelihood, significantly easing the derivation of the posterior distribution. Therefore, suppose that the four parameters are independent *a priori*, with

$$\alpha \sim \text{Beta}(\sigma_\alpha, \mu_\alpha),$$

$$\beta \sim \text{Beta}(\sigma_\beta, \mu_\beta),$$

$$\pi_0 \sim \text{Beta}(\sigma_{\pi_0}, \mu_{\pi_0}),$$

$$\pi_1 \sim \text{Beta}(\sigma_{\pi_1}, \mu_{\pi_1}),$$

giving a joint prior density

$$p(\alpha, \beta, \pi_0, \pi_1) = p_\alpha(\alpha)p_\beta(\beta) \prod_{i=0,1} p_{\pi_i}(\pi_i). \quad (2.9)$$

Since a posterior density is proportional to a likelihood multiplied by a prior, we have that the posterior density $p(\pi_0, \pi_1, \alpha, \beta | A_i, B_i, x_0, y_0, x_1, y_1)$ is proportional to

$$\begin{aligned} & \prod_{i=0,1} \left[\frac{N_i}{A_i! \dots (y_i - B_i)!} \pi_i^{A_i+B_i+\sigma_{\pi_i}-1} (1 - \pi_i)^{N_i-A_i-B_i+\mu_{\pi_i}-1} \right] \\ & \times \alpha^{A_0+A_1+\sigma_{\alpha}-1} (1 - \alpha)^{B_0+B_1+\mu_{\alpha}-1} \\ & \times \beta^{y_0+y_1-B_0-B_1+\sigma_{\beta}-1} (1 - \beta)^{x_0+x_1-A_0-A_1+\mu_{\beta}-1}. \end{aligned} \quad (2.10)$$

Note here that the latent data A_i and B_i , $i = 0, 1$ are not observed, hindering the use of (2.10) in calculating the marginal posterior densities of α , β , π_0 and π_1 . However, the inference is made possible by using a Gibbs sampler. This is a very useful technique for sampling from a p -dimensional distribution. Here is a brief review.

Suppose $q_{\mathbf{X}}$ is the joint density distribution function of a p -dimensional random variable $\mathbf{X} = (X_1, X_2, \dots, X_p)$, with the univariate conditional densities $q_{X_1|X_2, X_3, \dots, X_p}$, $q_{X_2|X_1, X_3, \dots, X_p}$, and so on. To implement the Gibbs sampler, we start with initial guesses of the X_i , say $X_1^{(0)}, X_2^{(0)}, \dots, X_p^{(0)}$ and simulate

$$\begin{aligned} & X_1^{(1)} | X_2^{(0)}, X_3^{(0)}, \dots, X_p^{(0)} \quad \text{from} \quad q_{X_1|X_2, X_3, \dots, X_p}, \\ & X_2^{(1)} | X_1^{(1)}, X_3^{(0)}, \dots, X_p^{(0)} \quad \text{from} \quad q_{X_2|X_1, X_3, \dots, X_p}, \\ & \quad \quad \quad \vdots \quad \quad \quad \vdots \\ & X_p^{(1)} | X_1^{(1)}, X_2^{(1)}, \dots, X_{p-1}^{(1)} \quad \text{from} \quad q_{X_p|X_1, X_2, \dots, X_{p-1}}. \end{aligned}$$

This is repeated k times, generating the sample $\mathbf{X}^{(k)} = (X_1^{(k)}, X_2^{(k)}, \dots, X_p^{(k)})$.

At each stage the conditional distribution uses the most recent values of all the other components of \mathbf{X} . It can be shown that, as $k \rightarrow \infty$, the density of the samples approaches $q_{\mathbf{X}}$. In practice, the convergence is usually quite rapid. Once the convergence has occurred, subsequent samples can be generated either by restarting the algorithm with the new guesses, or continuing the algorithm at the current value $\mathbf{X}^{(k)}$.

We can now use the Gibbs sampler to sample from the posterior distribution (2.10). We have the following conditional densities:

$$\alpha | A_0, B_0, A_1, B_1, \sigma_\alpha, \mu_\alpha \sim \text{Beta}(A_0 + A_1 + \sigma_\alpha, B_0 + B_1 + \mu_\alpha),$$

$$\beta | x_0, y_0, x_1, y_1, A_0, B_0, A_1, B_1, \sigma_\beta, \mu_\beta \sim \text{Beta}(y_0 + y_1 - B_0 - B_1 + \sigma_\beta, x_0 + x_1 - A_0 - A_1 + \mu_\beta),$$

$$\pi_i | A_i, B_i, x_i, y_i, \sigma_{\pi_i}, \mu_{\pi_i} \sim \text{Beta}(A_i + B_i + \sigma_{\pi_i}, x_i + y_i - A_i - B_i - \mu_{\pi_i}), i = 0, 1,$$

$$A_i | \pi_i, \alpha, \beta, x_i \sim \text{Binomial}(x_i, \frac{\pi_i \alpha}{\pi_i \alpha + (1 - \pi_i)(1 - \beta)}), i = 0, 1,$$

$$B_i | \pi_i, \alpha, \beta, y_i \sim \text{Binomial}(y_i, \frac{\pi_i (1 - \alpha)}{\pi_i (1 - \alpha) + (1 - \pi_i) \beta}), i = 0, 1.$$

Hence, conditional on knowing the exact values of α , β , π_0 and π_1 , we can easily sample from the posterior distributions of the latent variables A_i and B_i . Conversely, conditional on A_i and B_i , sampling from the posterior distributions of α , β , π_0 and π_1 , and consequently from ϕ , is straightforward.

2.3.1 Examples

Example 1 To illustrate the use of the Gibbs sampler, consider a scenario

in which the true value of sensitivity is $\alpha = 0.85$, specificity $\beta = 0.90$ and the prevalences of exposure in controls and cases is, respectively, $\pi_0 = 0.08$ and $\pi_1 = 0.12$. These values imply that $\phi = 1.57$ or $\log \phi = 0.45$. Data are simulated using the true values for three different sample sizes, namely for $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$, with respective simulated values of $(x_0, x_1) = (37, 41)$, $(x_0, x_1) = (128, 147)$ and $(x_0, x_1) = (495, 603)$. The prior for the prevalence of exposure in controls and the prevalence of exposure in cases was chosen to be $Beta(1, 1)$. No prior knowledge was assumed for π_0 and π_1 , since the motivation for doing the case-control study is to make inference about odds-ratio. Further, suppose that the researcher's guesses at the sensitivity and specificity are 0.83 and 0.91, respectively, and that these guesses are accurate to within ± 0.05 . The information of the form $\gamma \pm \delta$ could be translated into a Beta prior with mode at γ and 0.95 probability on the interval $\gamma \pm \delta$. This implies $Beta(183.50, 38.38)$ prior for sensitivity and $Beta(128.50, 13.61)$ prior for specificity. Five independent chains of the Gibbs sampler were run with different starting values. The output is shown in Figure 2.1, Figure 2.2 and Figure 2.3.

The plots shown in these figures clearly indicate drifting behaviour by the Gibbs sampler and a lack of adequate mixing, even after the first 10000 observations were discarded. For example, as the plots in Figure 2.1 indicate, convergence was not achieved for none of the parameters. This may not be surprising because, as Gelfand and Sahu (1999) note, drifting behaviour may arise when the Gibbs sampler is applied to nonidentifiable models. The drifting

behaviour also seems to be the case for intermediate and large sample sizes, however, it is not as pronounced. Even though more formal diagnostic methods were not used in assessing the convergence of the Gibbs sampler, we believe that the plots in Figure 2.1 through Figure 2.3 present strong enough evidence of the inadequacy of the Gibbs sampler in this case.

Example 2 Consider now a different scenario where the true value of sensitivity and specificity is $\alpha = \beta = 0.95$ and prevalence of exposure in controls and cases is $\pi_0 = 0.06$ and $\pi_1 = 0.15$, respectively. These values imply $\phi = 2.76$ or $\log \phi = 1.02$. Using these true values, we simulated data $(x_0, x_1) = (13, 44)$, $(x_0, x_1) = (68, 141)$ and $(x_0, x_1) = (340, 595)$ for $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$ respectively. As before, $Beta(1, 1)$ was used as prior for both the prevalence of exposure in controls and the prevalence of exposure in cases. As for the sensitivity and specificity, we will assume that our guess at both the sensitivity and specificity is 0.95, and that it is accurate to within ± 0.05 . This translates into $Beta(99.70, 6.19)$ prior for both the sensitivity and specificity. Five independent chains were run with different starting values. The output is shown in Figure 2.4, Figure 2.5 and Figure 2.6.

We observe a similar behaviour of the Gibbs sampler as in Example 1, even though our guesses at the true values of the sensitivity and specificity were exact. For instance, note the behaviour of the fourth chain in Figure 2.4. It exhibits a completely different behaviour than the other four chains in the sense that it drifts towards a completely different value in the parameter space.

In particular, the mean value of the sensitivity obtained from the fourth chain is 0.03, whereas the mean value obtained from the other four chains is 0.57, a very big discrepancy. Moreover, no point estimates were close to the true value of 0.95 for the sensitivity. The 95% HPD credible intervals of the sensitivity and the prevalence of exposure in cases for every sample size considered did not contain the true values.

2.4 The Metropolis-Hastings Algorithm

The behaviour observed above may not be surprising. Gelfand and Sahu (1999) have noted that drifting behaviour in the Gibbs sampler can arise if it is applied to nonidentifiable models. Therefore, the extension of the methodology of Joseph, Gyorkos and Coupal (1995) to the case-control setting does not seem to be appropriate. Instead, we use the the Metropolis-Hastings algorithm in the reparameterization (2.3). This reparameterization separates the identifiable parameters (θ_0, θ_1) from the nonidentifiable parameters (α, β) . In this new parameterization, (2.9) becomes

$$\begin{aligned} p(\alpha, \beta, \theta_0, \theta_1) &= \frac{1}{(\alpha + \beta - 1)^2} p_\alpha(\alpha) p_\beta(\beta) \\ &\times I_{(0,1)}\left(\frac{\theta_0 + \beta - 1}{\alpha + \beta - 1}\right) I_{(0,1)}\left(\frac{\theta_1 + \beta - 1}{\alpha + \beta - 1}\right), \end{aligned} \quad (2.11)$$

since the Jacobian is given by

$$J(\theta_0, \theta_1) = \begin{vmatrix} \frac{\partial \pi_0}{\partial \theta_0} & \frac{\partial \pi_1}{\partial \theta_0} \\ \frac{\partial \pi_0}{\partial \theta_1} & \frac{\partial \pi_1}{\partial \theta_1} \end{vmatrix} = \frac{1}{(\alpha + \beta - 1)^2}.$$

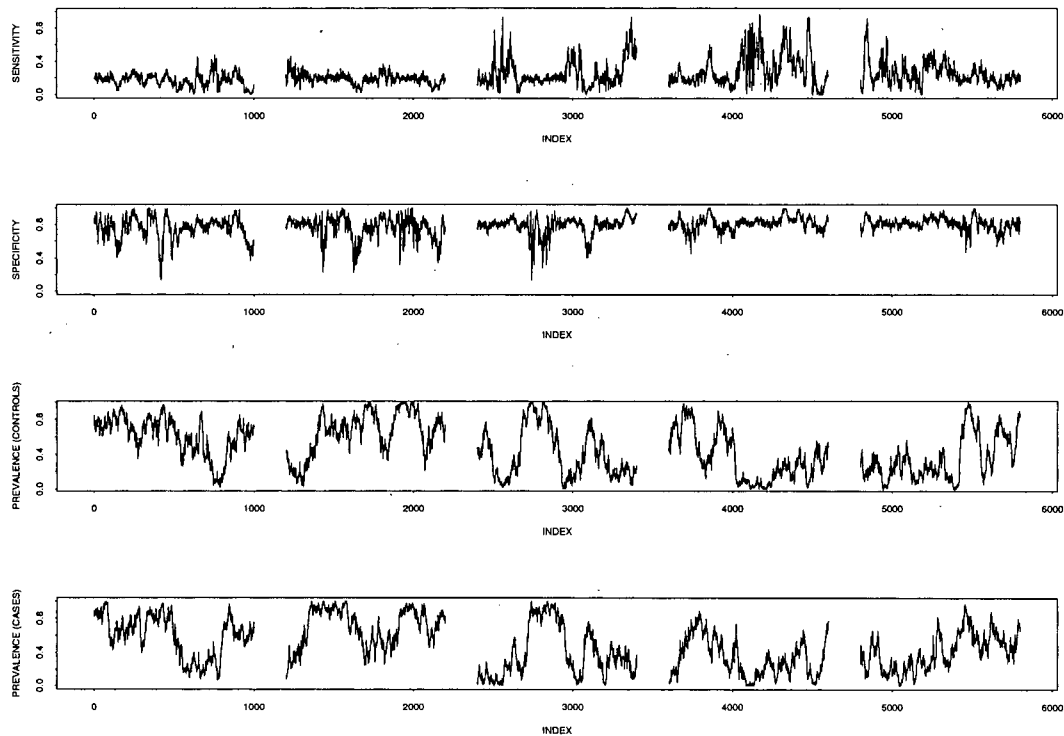


Figure 2.1: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 200$.

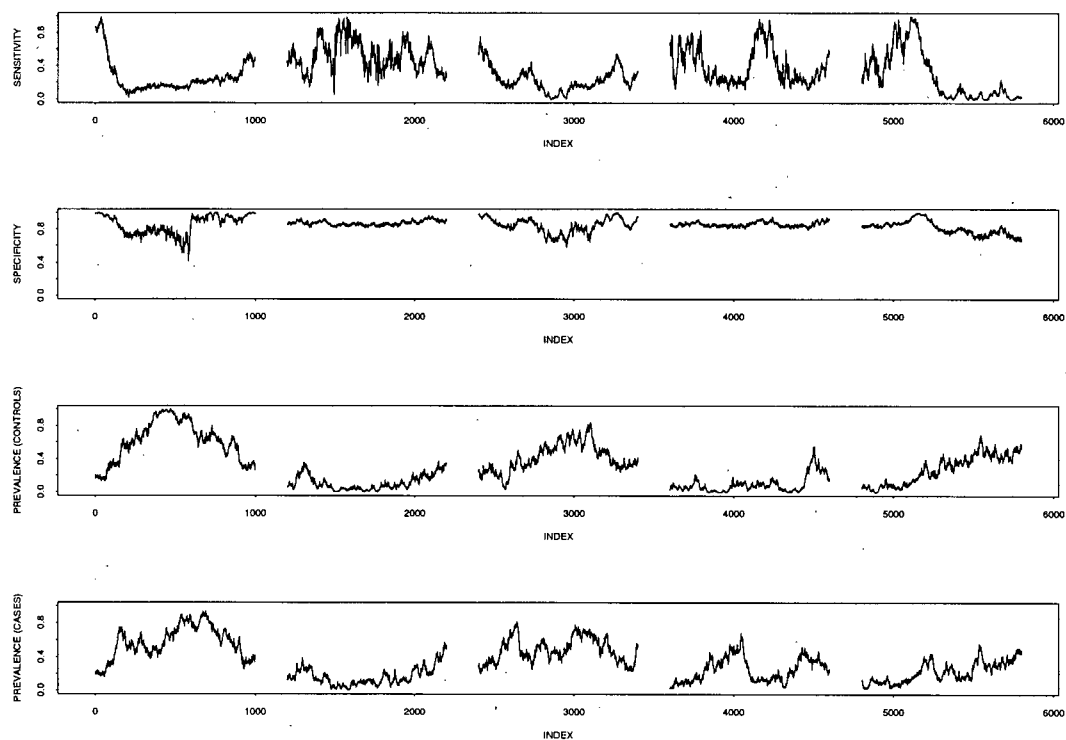


Figure 2.2: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 800$.

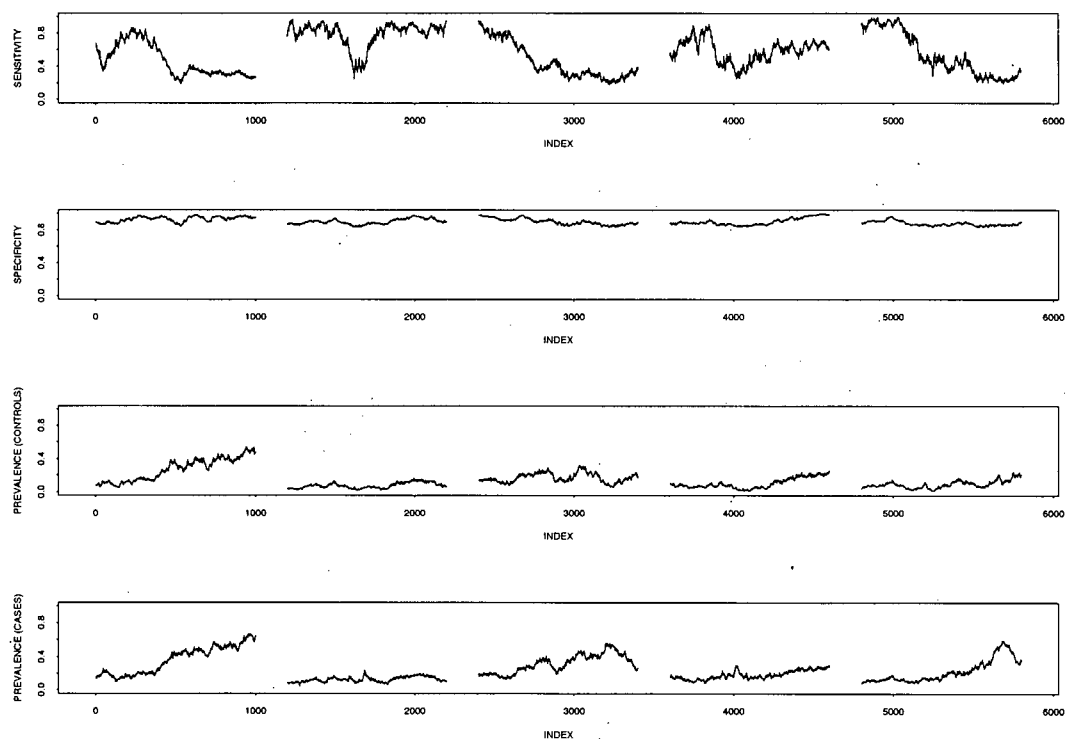


Figure 2.3: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 3200$.

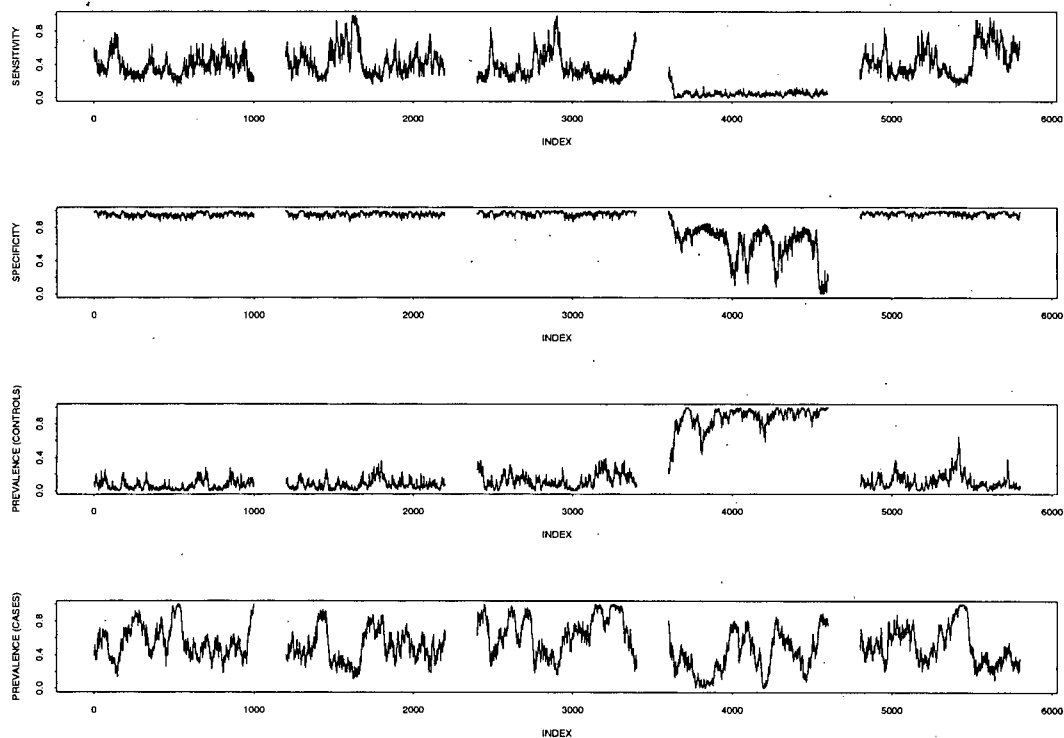


Figure 2.4: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 200$.

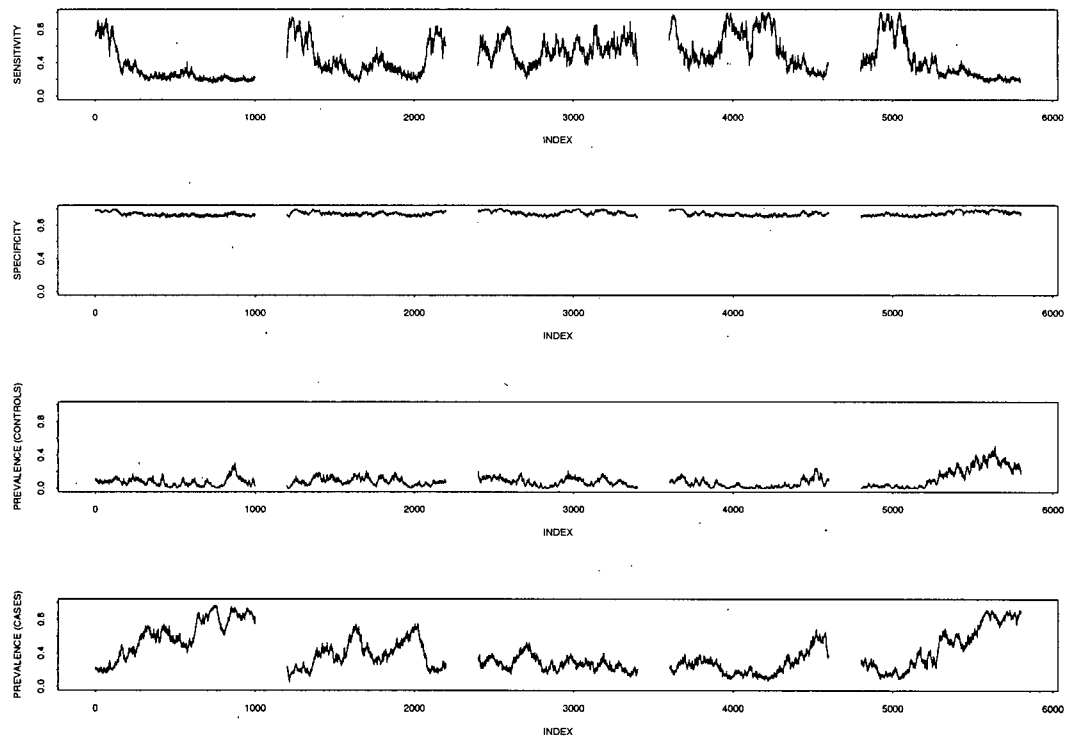


Figure 2.5: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 800$.

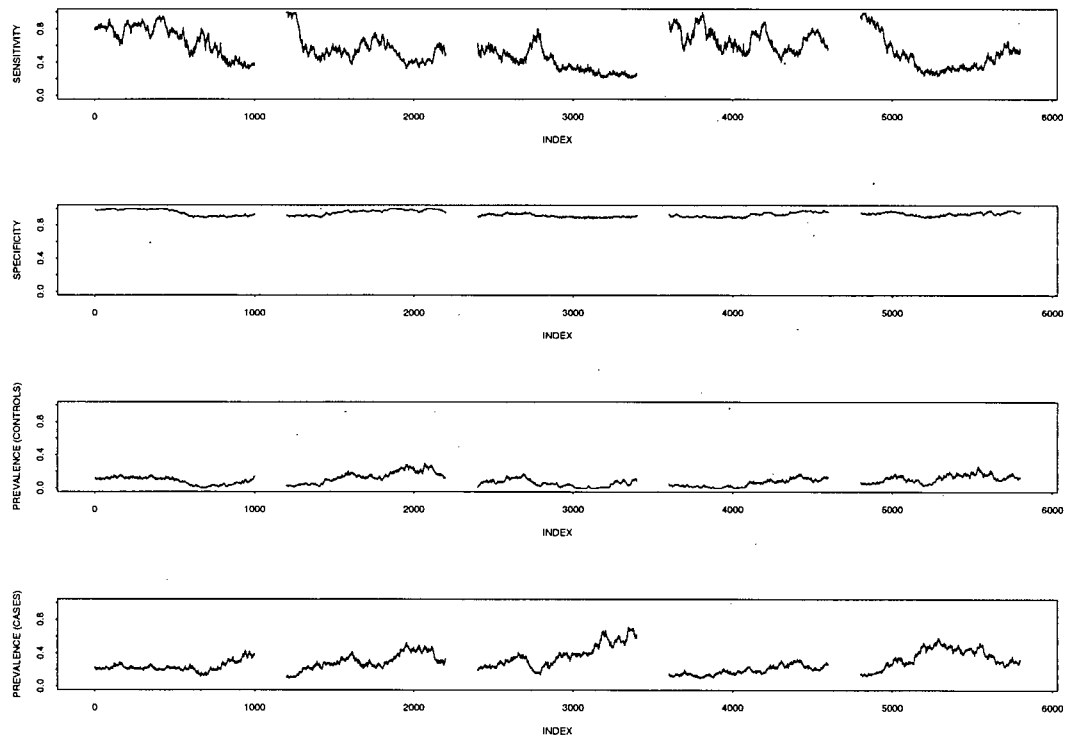


Figure 2.6: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivity, specificity, prevalence of exposure in controls and prevalence of exposure in cases in Example 2. Data was simulated for $N_0 = N_1 = 3200$.

Therefore, the joint posterior distribution of $(\alpha, \beta, \theta_0, \theta_1)$ given the data is

$$\begin{aligned}
p(\alpha, \beta, \theta_0, \theta_1 | x_0, x_1) &= \frac{p(x_0, x_1 | \theta_0, \theta_1, \alpha, \beta) p(\alpha, \beta, \theta_0, \theta_1)}{p(x_0, x_1)} \\
&= \frac{p(x_0, x_1 | \theta_0, \theta_1) p(\alpha, \beta, \theta_0, \theta_1)}{p(x_0, x_1)} \\
&= \left\{ \frac{p(x_0, x_1 | \theta_0, \theta_1) p(\theta_0, \theta_1)}{p(x_0, x_1)} \right\} p(\alpha, \beta, | \theta_0, \theta_1) \\
&= p(\theta_0, \theta_1 | x_0, x_1) p(\alpha, \beta, | \theta_0, \theta_1). \tag{2.12}
\end{aligned}$$

Here, the prior and posterior conditional distribution of $\alpha, \beta | \theta_0, \theta_1$ are identical, and hence, according to Dawid (1979), (α, β) are nonidentifiable parameters. However, as a result of learning about (θ_0, θ_1) , the marginal prior and posterior distributions of the sensitivity and specificity are not equal. Thus, one could learn about (α, β) indirectly, through the updating of (θ_0, θ_1) .

2.4.1 Large Sample Case

Let us examine (2.12) further. Since $p(\theta_0, \theta_1 | x_0, x_1)$ and $p(\theta_0, \theta_1)$ constitute a regular model and prior, the posterior distribution of $\theta_0, \theta_1 | x_0, x_1$ will converge to a point mass at the true value of (θ_0, θ_1) , as $N_0 \rightarrow \infty$ and $N_1 \rightarrow \infty$. The uncertainty in $p(\alpha, \beta, | \theta_0, \theta_1)$ remains unchanged by any amount of data. Note that we can express it as

$$\begin{aligned}
p(\alpha, \beta | \theta_0, \theta_1) &\propto p(\theta_0, \theta_1 | \alpha, \beta) p(\theta_0, \theta_1) \\
&\propto p(\theta_0 | \theta_1, \alpha, \beta) p(\theta_1 | \alpha, \beta) p_\alpha(\alpha) p_\beta(\beta) \\
&\propto p(\theta_0 | \alpha, \beta) p(\theta_1 | \alpha, \beta) p_\alpha(\alpha) p_\beta(\beta)
\end{aligned}$$

$$\propto \frac{1}{(\alpha + \beta - 1)^2} I_{A(\theta_0, \theta_1)}(\alpha, \beta) p_\alpha(\alpha) p_\beta(\beta), \quad (2.13)$$

where

$$\begin{aligned} A(\theta_0, \theta_1) &= \{(\alpha, \beta) : \bar{\theta} < \alpha < 1, 1 - \underline{\theta} < \beta < 1\} \\ &\cup \{(\alpha, \beta) : 0 < \alpha < \underline{\theta}, 0 < \beta < 1 - \bar{\theta}\}, \end{aligned}$$

and $\bar{\theta} = \max\{\theta_0, \theta_1\}$ and $\underline{\theta} = \min\{\theta_0, \theta_1\}$. This region is shown in Figure 2.7.

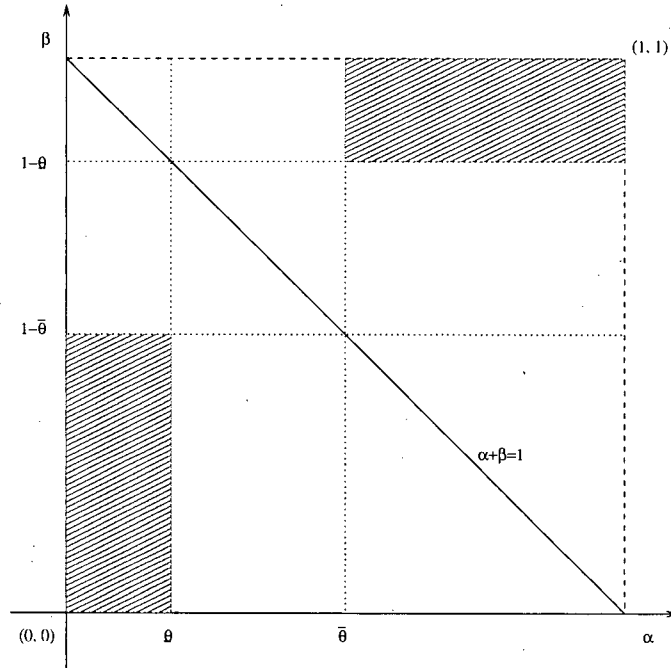


Figure 2.7: The support of $p(\alpha, \beta | \theta_0, \theta_1)$. The shaded rectangles comprise $A(\theta_0, \theta_1)$.

From (2.13) we see that $p(\alpha, \beta)$ differs from $p(\alpha, \beta | \theta_0, \theta_1)$ in two respects. First, the support of $p(\alpha, \beta)$ is the whole unit square, whereas the support of

$p(\alpha, \beta | \theta_0, \theta_1)$ is $A(\theta_0, \theta_1)$. Blettner and Wahrendorf (1984) observed this in the context of finite samples. They describe how to calculate the range of the possible underlying true effects, measured by odds-ratio, and of the possible misclassification probabilities, given the observed data. Second, the density $p(\alpha, \beta | \theta_0, \theta_1)$ has $(\alpha + \beta - 1)^{-2}$ as a factor. It is, of course, infinite along the line $\alpha = 1 - \beta$. The rectangles that make up the support $A(\theta_0, \theta_1)$ are on either side of this line, implying that (2.13) is bounded. In particular,

$$\frac{1}{(\alpha + \beta - 1)^2} I_{A(\theta_0, \theta_1)}(\alpha, \beta) p_\alpha(\alpha) p_\beta(\beta) \leq \frac{1}{(\theta_1 - \theta_0)^2} I_{A(\theta_0, \theta_1)}(\alpha, \beta) p_\alpha(\alpha) p_\beta(\beta). \quad (2.14)$$

Therefore, we can sample from (2.13) using the “acceptance/rejection algorithm”. To sample from a random variable X , this algorithm makes use of samples of another random variable, say Y , whose probability density function g_Y is similar to the probability density function of X , p_X . The random variable Y is chosen so that we can easily generate samples of it and so that its density g_Y can be scaled to majorize p_X , using some constant c ; that is, so that $cg_Y(x) \geq p_X(x)$, for all x . The density g_Y is called the “majorizing density” and cg_Y is called the “majorizing function”. The closer $cg_Y(x)$ is to $p_X(x)$, the faster the acceptance/rejection algorithm will be. In our case, the majorizing density is $p_\alpha(\alpha)p_\beta(\beta)$ truncated to the interval $A(\theta_0, \theta_1)$. For a more detailed description, see Gentle (1998).

However, we have not found the acceptance/rejection algorithm to be very efficient, and instead used the “Metropolis-Hastings algorithm”. The Metropolis-Hastings algorithm helps when it is difficult to sample from the

target distribution, say $\pi(x)$, but there is an easy way to sample from another, say $q(x, y)$, called the “candidate-generating density”. In our case, the candidate-generating density is $p_\alpha(\alpha)p_\beta(\beta)$ truncated to the interval $A(\theta_0, \theta_1)$, same as the majorizing density above. The Metropolis-Hastings algorithm proceeds by constructing a Markov Chain which has $\pi(x)$ as its stationary distribution. We present a brief illustration of how the Metropolis-Hastings algorithm works. For a complete description, see Chib and Greenberg (1995) or Gentle (1998).

Let X be an irreducible, recurrent Markov Chain, with the stationary distribution $\pi(x)$, and let x be its current state. The Metropolis-Hastings algorithm proceeds by sampling a candidate value y from $q(x, y)$, and choosing the next state of the Markov Chain to be either the candidate value y or the current state x , depending on the following acceptance probability

$$a(x, y) = \begin{cases} \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\} & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{otherwise} \end{cases} \quad (2.15)$$

The process is repeated k times. Of course, the draws are regarded as a sample from the target density $\pi(x)$ only as k gets large.

The above algorithm has a drawback in comparison with the Gibbs sampler: the degree of serial correlation is increased due to the possibility of rejection of a move. If $q(x, y)$ is such that the probability of a move is very small, the chain will be characterized by long sequences of repeated outcomes. However, a very convenient feature of the Metropolis-Hastings algorithm, apart from the free selection of the candidate-generating density $q(x, y)$, is that we only need to know $\pi(x)$ up to a normalizing constant (as it cancels out in the

ratio $\pi(y)/\pi(x)$.

Different choices of $q(x, y)$ give rise to different variants of the Metropolis-Hastings algorithm. If $q(x, y)$ is symmetric, i.e., if $q(x, y) = q(y, x)$, the probability of a move reduces to $\min(\pi(y)/\pi(x), 1)$. This is known as the “Metropolis algorithm”. Another family of the candidate-generating densities is given by $q(x, y) = \tilde{q}(y - x)$, characterizing the “random walk” chain. The candidate value y is given by $y = x + z$, where Z is a random variable drawn from $\tilde{q}(z)$.

Our choice of $q(x, y)$ yields what is known as the “independence” chain. It arises when the candidate-generating density is not a function of the current state, i.e., when $q(x, y) = \hat{q}(y)$, for some density $\hat{q}(\cdot)$. Other choices of the candidate-generating density are possible.

2.4.2 Finite Sample Case

In a finite sample case, we have demonstrated that the Gibbs sampler does not work very well. Instead, we try to sample from (2.12) using the Metropolis-Hastings algorithm. Note that (2.12) can be expressed as

$$p(\alpha, \beta, \theta_0, \theta_1 | x_0, x_1) \propto \theta_0^{x_0} (1 - \theta_0)^{N_0 - x_0} \theta_1^{x_1} (1 - \theta_1)^{N_1 - x_1} \\ \times \frac{1}{(\alpha + \beta - 1)^2} I_{A(\theta_0, \theta_1)}(\alpha, \beta) p_\alpha(\alpha) p_\beta(\beta). \quad (2.16)$$

The candidate generating density that approximates (2.16) from which we can sample to drive the Metropolis-Hastings algorithm is

$$f(\alpha, \beta, \theta_0, \theta_1) = \frac{1}{k(\theta_0, \theta_1)} f_0(\theta_0) f_1(\theta_1) I_{A(\theta_0, \theta_1)}(\alpha, \beta) p_\alpha(\alpha) p_\beta(\beta). \quad (2.17)$$

Here, $k(\theta_0, \theta_1) = \Pr\{(\alpha, \beta) \in A(\theta_0, \theta_1)\}$, so that $f(\alpha, \beta|\theta_0, \theta_1)$ is the prior $p(\alpha, \beta)$ truncated to $A(\theta_0, \theta_1)$. Furthermore, f_i is the $Beta(x_i + 1, N_i - x_i + 1)$ density, so that $p(\alpha, \beta|\theta_0, \theta_1)$ in (2.12) is approximated by $f(\theta_0, \theta_1)$ by replacing the intractable prior for (θ_0, θ_1) by a uniform prior on the unit square. Therefore, we can sample from (2.17) by first sampling from $f(\theta_0, \theta_1)$ and then from $f(\alpha, \beta|\theta_0, \theta_1)$.

We now present three examples illustrating the use of Metropolis-Hastings algorithm.

2.4.3 Examples

Example 1 Consider a scenario in which the true value of sensitivity and specificity is $\alpha = \beta = 0.84$, and the prevalences of exposure in controls and cases is, respectively, $\pi_0 = 0.061$ and $\pi_1 = 0.15$, implying that $\theta_0 = 0.20148$, $\theta_1 = 0.262$ and $\log \phi = 1.00$. Furthermore, suppose that our best guesses at the sensitivity and specificity are $(\alpha', \beta') = (0.81, 0.81)$ and that these guesses are accurate to within ± 0.05 . This leads to a $Beta(194.00, 46.27)$ prior for both α and β . Under the guessed values, $\log \phi' = 1.94$. Data are simulated under this scenario for three different sample sizes, namely for $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$, with respective simulated values of $(x_0, x_1) = (41, 54)$, $(x_0, x_1) = (162, 214)$ and $(x_0, x_1) = (661, 861)$. Three different posterior distributions were used to generate the samples of $\log \phi$ for each of the data sets. The first, named “miscorrected” (M-COR) posterior, originates from assuming $(\alpha, \beta) = (\alpha', \beta')$. The second, named “uncertainty-

corrected" (U-COR) posterior originates from assuming a prior distribution on α and β . The third, named "exactly-corrected" (E-COR) posterior arises from the knowledge of the exact values of the sensitivity and specificity. The sampling from M-COR and E-COR proceeds as described in section 2.2, whereas from U-COR as per section 2.4.2. Furthermore, we consider the large sample case where the M-COR and E-COR concentrate at ϕ' and ϕ , respectively, and the sampling from U-COR proceeds as per section 2.4.1.

All posterior distributions are shown in Figure 2.8. There is very little difference between the three posteriors for $N_0 = N_1 = 200$. At the intermediate sample size, M-COR and U-COR posteriors are similar, with E-COR posterior being more peaked. A greater difference between M-COR and U-COR posteriors appears for the large sample size. The U-COR posterior covers both the true value of $\log \phi$ and the wrong value $\log \phi'$, whereas the M-COR posterior misses the true value and is drawn towards $\log \phi'$. Even in the large- n limit, the U-COR posterior is quite wide, covering both values. However, this should still be preferable to the M-COR posterior which, asymptotically, concentrates at the wrong answer. Convergence is very slow in the present scenario.

To assess the coverage of the credible intervals under each posterior, we simulate forty data sets under the same scenario for the large sample size $N_0 = N_1 = 3200$ and compute 80% HPD credible intervals for $\log \phi$ under the three posteriors. These intervals appear in Figure 2.9. The empirical coverage rates for the intervals are 22%, 90%, 90% and the average interval widths are 1.76, 1.64, 0.56 for the M-COR, U-COR and E-COR intervals

respectively. Note the very poor empirical coverage exhibited by the M-COR credible intervals as many of the intervals are drawn toward $\log \phi'$. On the other hand, the U-COR intervals show overcoverage. However, the cost of admitting uncertainty seems to be high, as the average length of the U-COR intervals is 1.64 on the log scale.

As mentioned in section 2.4.1, the prior and the posterior distribution for (α, β) differ in two respects, one of which is the support. Figure 2.10 shows the effects of truncation to $A(\theta_0, \theta_1)$. We see that this effect is most evident for the case $N_0 = N_1 = \infty$. Furthermore, note that the truncation affects β but not α . This will usually be true for small exposure prevalences. When π_0 and π_1 are small, $1 - \theta_0$ and $1 - \theta_1$ will lie in the region where draws from $p_\beta(\beta)$ are likely, whereas θ_0 and θ_1 will be in the region where draws from $p_\alpha(\alpha)$ are not likely.

Example 2 Consider now a scenario in which the true values of the sensitivity and specificity are $\alpha = 0.84$ and $\beta = 0.78$, and all other quantities are the same as in Example 1. This implies $\theta_0 = 0.25782$, $\theta_1 = 0.313$ and $\log \phi' = 0.70$. Under these conditions, we simulate data $(x_0, x_1) = (55, 67)$, $(x_0, x_1) = (190, 256)$ and $(x_0, x_1) = (830, 966)$ for $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$ respectively. Posterior distributions for these data are shown in Figure 2.11. For the small sample size, there is very little difference between the M-COR and U-COR posteriors. For the intermediate and large sample sizes, the U-COR posterior performs better in the sense that it is closer to the E-COR posterior. The M-COR posterior again misses the

true value of the log odds-ratio.

Forty data sets are again considered to assess the coverage of the credible intervals for the large sample size under each posterior. The 80% HPD credible intervals are shown in Figure 2.12. The empirical coverage rates for the intervals are 37%, 87%, 82% and the average interval widths are 0.40, 0.77, 0.67 for the M-COR, U-COR and E-COR intervals respectively. Again, we see the undercoverage of the M-COR intervals, and the slight overcoverage of the U-COR intervals. In terms of the average widths of the intervals, the cost of admitting uncertainty in this example does not seem to be as high as in Example 1.

Example 3 Finally, suppose that our best guesses at the sensitivity and specificity are $(\alpha', \beta') = (1, 1)$, analogous to no misclassification. Further, suppose that we feel that there could be small misclassification, so that $Beta(58.41, 1)$ was assigned as a prior to both α and β . This prior density is a strictly increasing function with finite maximum at 1, and 95% probability assigned to the interval $[0.95, 1]$. Suppose that the prevalences are the same as in Example 1, but the sensitivity and specificity are $\alpha = \beta = 0.97$. These values imply $\theta_0 = 0.08734$, $\theta_1 = 0.171$ and $\log \phi' = 0.77$. Again, we simulate data under this scenario for $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$, with respective simulated values being $(x_0, x_1) = (23, 31)$, $(x_0, x_1) = (63, 140)$ and $(x_0, x_1) = (285, 575)$. All posteriors appear in Figure 2.13. Again, for the small sample size, there is very little difference between the three posteriors. However, in the case of the intermediate and the large sample size, the

U-COR posterior is closer to the E-COR posterior than is M-COR posterior, showing once more the value of admitting uncertainty about the sensitivity and specificity.

Again, we simulate 40 data sets under this scenario for $N_0 = N_1 = 3200$. The resulting 80% HPD credible intervals are shown in Figure 2.14. Similar to Examples 1 and 2, the empirical coverage of the M-COR intervals is very low (3%), while 85% of the U-COR intervals cover the true value. The average interval widths are 0.20, 0.41, 0.28 for the M-COR, U-COR and E-COR intervals respectively. Hence, the cost of admitting uncertainty is not nearly as severe as in Example 1.

2.5 Discussion

So far, we have focused our attention to a case-control setting where one imperfect classification scheme is used to assess the exposure status. The odds-ratio estimates obtained from the observed data, if not corrected for misclassification, can be very imprecise. Relatively straightforward correction methods are available if the classification probabilities are known. However, the odds-ratio estimates which are corrected using slightly inaccurate classification probabilities may still be quite erroneous. In particular, we have demonstrated that even arbitrarily small differences between the true and assumed classification probabilities can, asymptotically, lead to arbitrarily large difference between the actual odds-ratio and the odds-ratio obtained using the assumed classification probabilities. By admitting uncertainty about the classification probabil-

ities, this problem seems to be alleviated. Two computational approaches to this analysis were suggested: the Gibbs sampler and the Metropolis-Hastings algorithm. The Gibbs sampler did not seem to work particularly well. This was not surprising since, as Gelfand and Sahu (1999) noted, drifting behaviour of the Gibbs sampler may occur if it is applied to a nonidentifiable model. On the other hand, the Metropolis-Hastings algorithm applied to the model that separates the identifiable and the nonidentifiable part seemed to perform well.

Despite nonidentifiability, there was considerable learning about the odds-ratio from the data. The U-COR posterior was quite concentrated relative to the prior. However, the cost of admitting uncertainty is evident in the increase in variability of the U-COR posterior relative to the E-COR posterior. As a result, the U-COR credible intervals were twice as wide as the E-COR credible intervals. On the other hand, there seemed to be an improvement of the U-COR posterior over the M-COR posterior. This improvement appeared to arise from marginal learning about the specificity but not about the sensitivity.

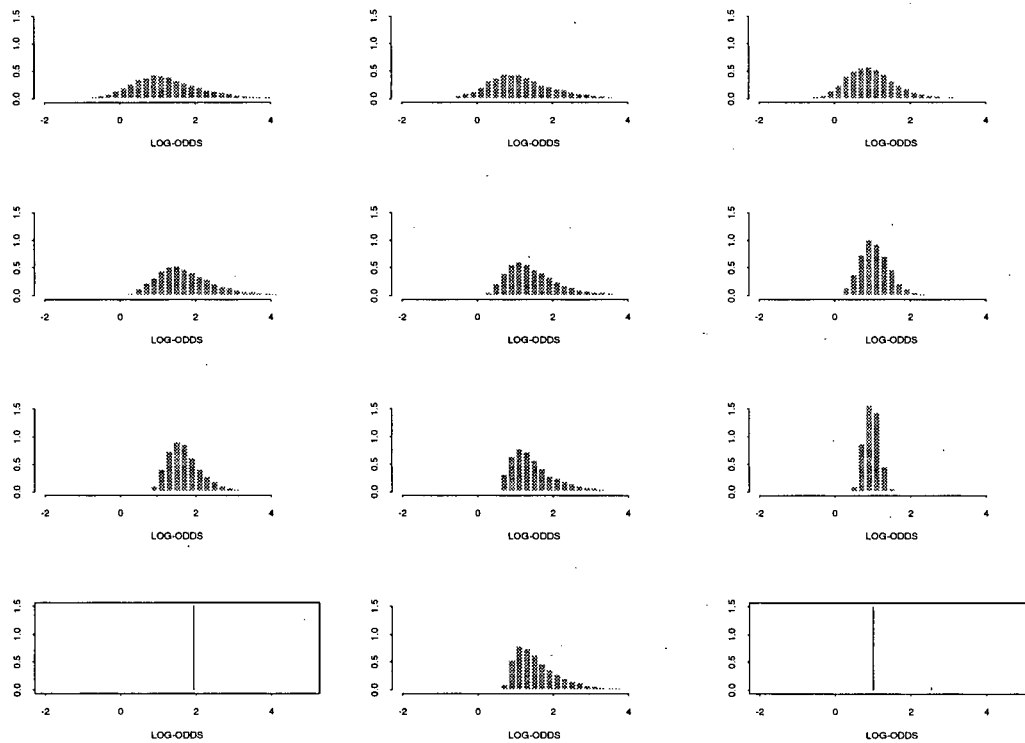


Figure 2.8: Posterior distribution of $\log \phi$ in Example 1. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$.

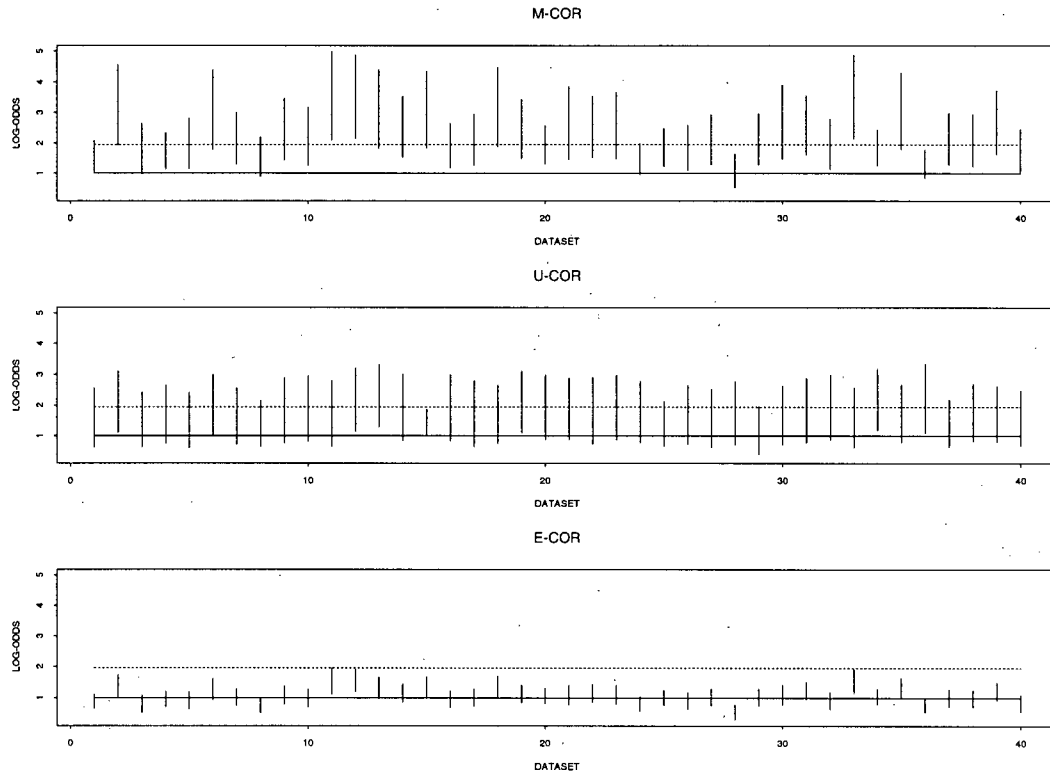
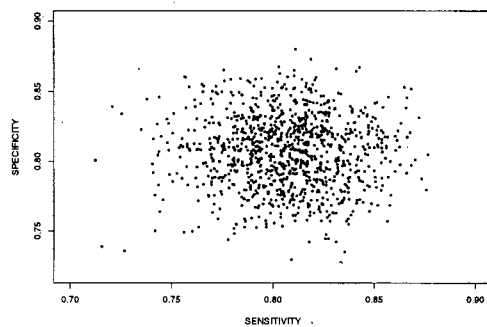
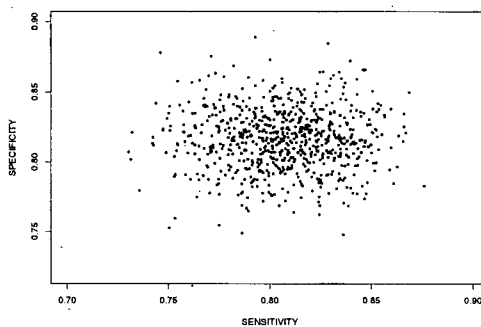


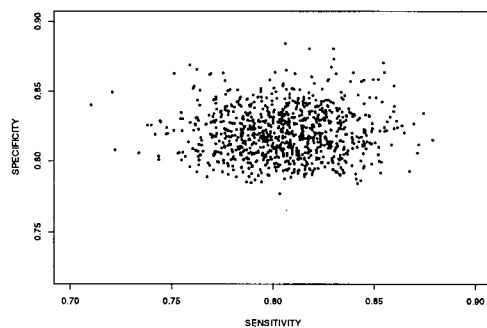
Figure 2.9: 80% highest posterior density credible intervals for $\log \phi$ in Example 1. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.



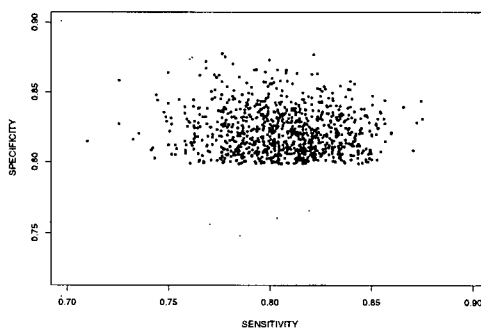
(a) Prior



(b) Sample size: $N_0 = N_1 = 200$



(c) Sample size: $N_0 = N_1 = 3200$



(d) Sample size: $N_0 = N_1 = \infty$

Figure 2.10: Prior and posterior samples of the sensitivity and specificity for the datasets in Example 1.

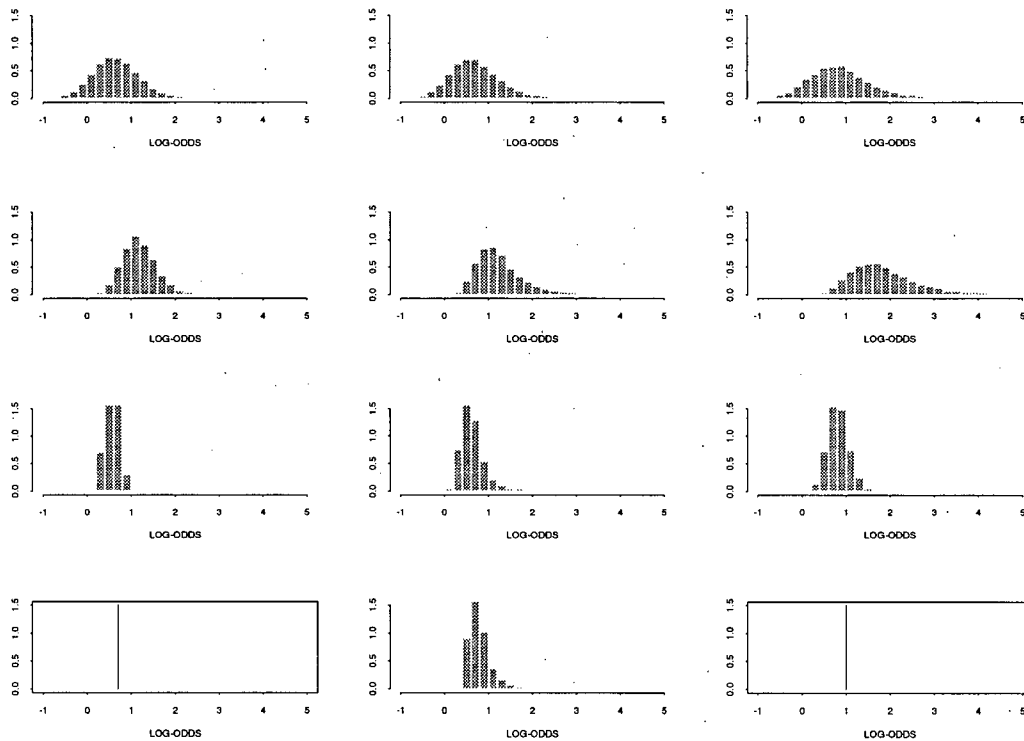


Figure 2.11: Posterior distribution of $\log \phi$ in Example 2. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$.

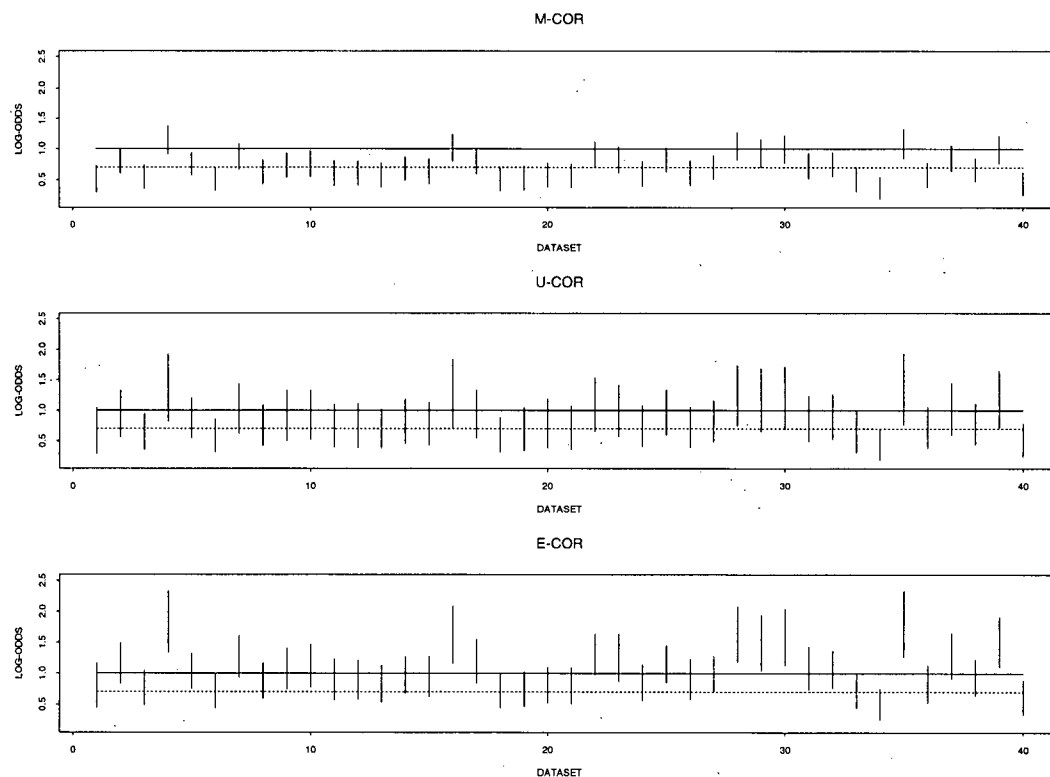


Figure 2.12: 80% highest posterior density credible intervals for $\log \phi$ in Example 2. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.

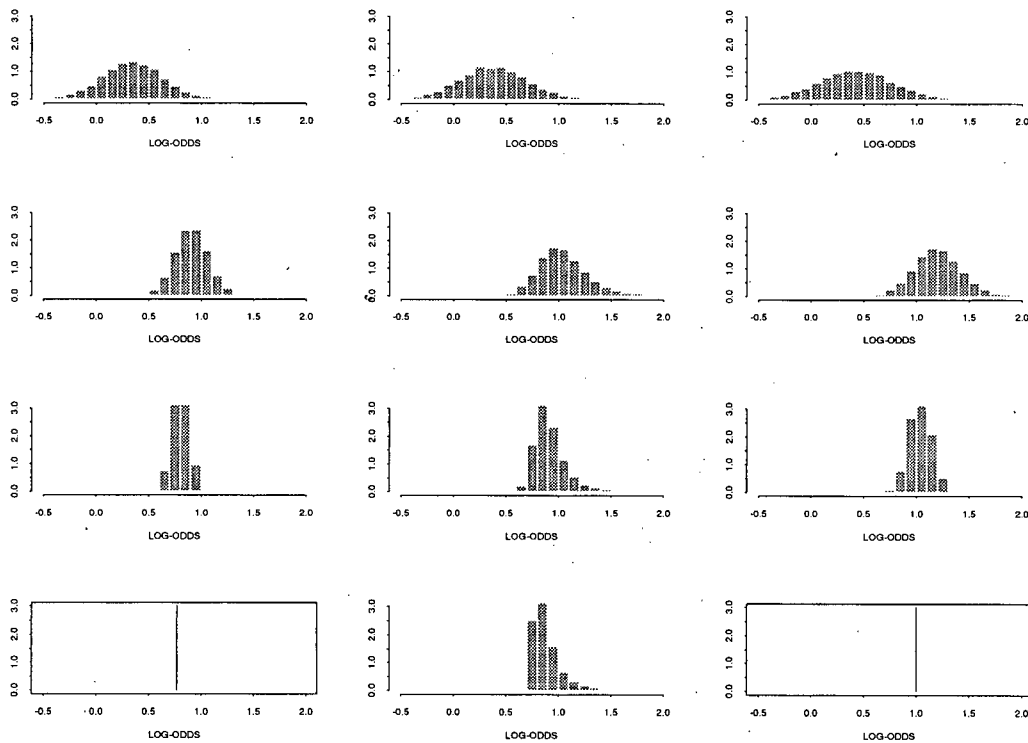


Figure 2.13: Posterior distribution of $\log \phi$ in Example 3. The first column gives the M-COR posterior, the second column gives the U-COR posterior, and the third column gives the E-COR posterior. The rows correspond to sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, $N_0 = N_1 = 3200$ and $N_0 = N_1 = \infty$.

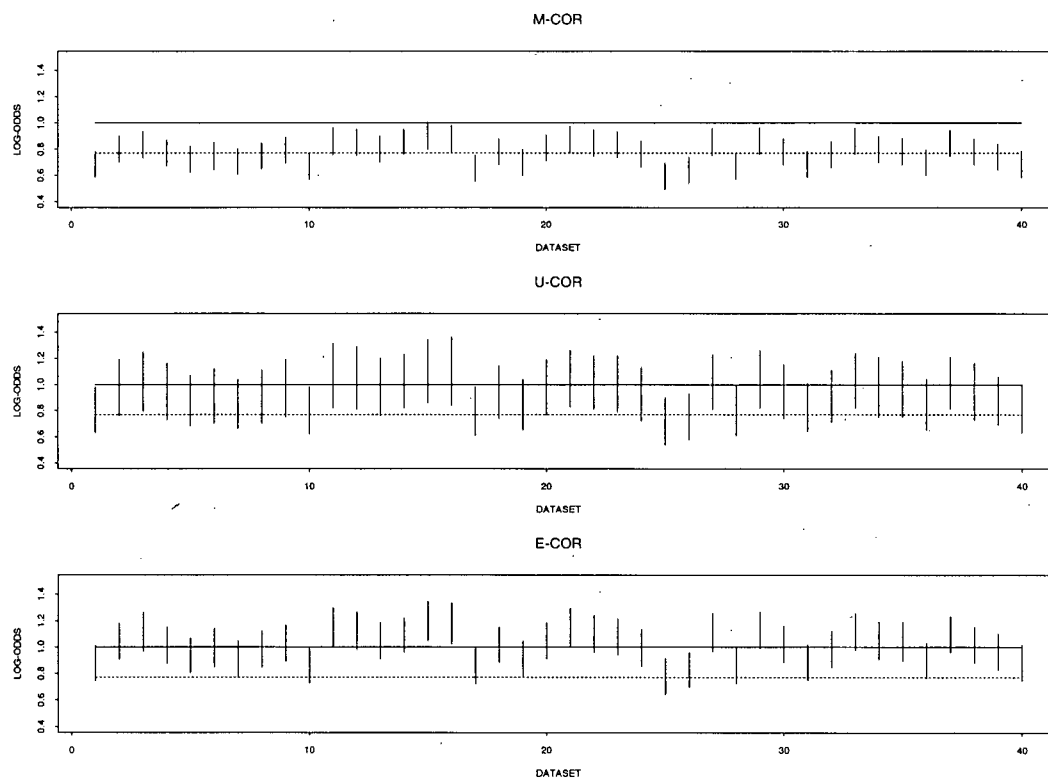


Figure 2.14: 80% highest posterior density credible intervals for $\log \phi$ in Example 3. The solid vertical lines represent credible intervals for forty data sets with sample sizes $N_0 = N_1 = 3200$, the solid horizontal line indicates the true value of $\log \phi$ and the dashed horizontal line indicates $\log \phi'$. The first panel gives the M-COR intervals, the second panel gives the U-COR intervals, and the third panel gives the E-COR intervals.

Chapter 3

Two Test Case

In the previous chapter we examined a case-control setting where exposure is measured with one imperfect test. We now turn our attention to the situation where the exposure status is determined by two imperfect classification schemes. We will illustrate the existing methods to correct the odds-ratio estimates, and examine our method via a simulation study and an analysis of data on the sudden infant death syndrome.

3.1 Data and Setup

As in chapter 2 we define relevant parameters and introduce the general setup. Consider the data from a case-control study in which each subject has an underlying true, but unobserved exposure (E), coded as 1 for exposed and 0 for unexposed. This exposure is assessed by applying two measures or tests (say, T_1 and T_2) to each subject. These tests are coded as 1 for positive and

0 for negative outcome. We will assume that the classification procedures misclassify the subjects nondifferentially and that the disease status is known exactly, without error. When two tests are used to determine the exposure status, the data usually come in the following form (Table 3.1).

Controls				Cases					
		Test 2				Test 2			
		+	-			+	-		
Test 1	+	a_0	b_0	$a_0 + b_0$	Test 1	+	a_1	b_1	$a_1 + b_1$
	-	c_0	d_0	$c_0 + d_0$		-	c_1	d_1	$c_1 + d_1$
				N_0					N_1

Table 3.1: Distribution of subjects in a case-control study when two tests are used to assess the exposure status. Symbol + denotes exposed and - denotes unexposed.

We define the following parameters to model the probabilities of the different possible outcomes in controls and cases:

$$\alpha_1 = \Pr(T_1 = 1|E = 1),$$

$$\alpha_2 = \Pr(T_2 = 1|E = 1),$$

$$\beta_1 = \Pr(T_1 = 0|E = 0),$$

$$\beta_2 = \Pr(T_2 = 0|E = 0),$$

with the exposure prevalences π_0 and π_1 and the odds-ratio ϕ defined as before.

Here, α_1 and α_2 are the sensitivities of test 1 and test 2, respectively, and β_1 and β_2 are the specificities of test 1 and test 2, respectively.

3.2 Correction Methods

In this section we illustrate three different methods to correct the odds-ratio estimates using dual measurements. The first method uses only the concordant data (i.e., the data for which two measurements are in agreement). The other two methods use the EM algorithm to obtain the estimates. This algorithm is a useful way of obtaining maximum likelihood estimates when some data are missing. Each iteration of the EM algorithm involves two steps: the expectation step (E-step) and the maximization step (M-step). For precise definitions of these steps and the full description of the EM algorithm, see Dempster, Laird and Rubin (1977).

3.2.1 The Marshall and Graham Method

Marshall and Graham (1984) consider the problem of exposure to a risk factor, where the “true” exposure is unknown. They proposed a simple way to decrease the bias in the odds-ratio estimates caused by misclassification, using two independent imperfect tests to gather information on exposure status. The method is based on restricting the analysis to data for which two independent assessments of exposure are concordant, either positive or negative. Subjects for which both tests are positive are considered as exposed, and those for which both tests are negative are considered as unexposed.

They have demonstrated that the use of two tests in this fashion can provide a less biased estimate of the odds-ratio. Even if the second test is less accurate than the first, the contrast between subjects about whom there is test agreement yielded a better approximation of the odds-ratio than if only the more accurate report were used.

There are, of course, obvious drawbacks of this method. The Marshall-Graham procedure uses only a subset of the data, the concordant observations. Clearly, by using only a part of the data, the statistical efficiency of the estimate is decreased. Furthermore, this procedure does not completely remove the bias caused by misclassification. It is necessary that both the sensitivity and specificity of both assessment procedures be relatively high, in order to closely approximate the relative risk. For example, if the sensitivity and specificity are about 0.9, the prevalence of exposure in cases is 0.3 and the prevalence of exposure in controls is about 0.097, the true odds-ratio of 4.0 will be approximated by about 3.7.

3.2.2 The Drews, Flanders and Kosinski Method

Drews, Flanders and Kosinski (1993) examine methods to use two classification schemes to improve the odds ratio estimates in case-control studies. Their assumptions are differ slightly from ours in that they do not assume that errors in one test are independent of errors in the second test. This is reflected in the following parameters used by these authors, in addition to the sensitivity and specificity of test 1, to model the probabilities of the different possible

outcomes in cases and controls:

$$\begin{aligned}\alpha_2 &= Pr(T_2 = 1|E = 1, T_1 = 1), \\ \alpha_2\mu_1 &= Pr(T_2 = 1|E = 1, T_1 = 0), \\ \beta_2 &= Pr(T_2 = 0|E = 0, T_1 = 1), \\ \beta_2\mu_0 &= Pr(T_2 = 0|E = 0, T_1 = 0).\end{aligned}$$

Under this model, μ_1 and μ_0 represent deviations from a model in which, given true exposure status, errors in one test are independent of errors in a second test. Note also that if the misclassification probabilities of T_2 are independent of the misclassification probabilities of T_1 (that is, $\mu_1 = \mu_0 = 1$), α_2 and β_2 are not the sensitivity and specificity of T_2 . Since there are two tests applied to two populations, there are six degrees of freedom to estimate six parameters. Hence, the likelihood is identifiable, provided that we treat μ_1 and μ_0 as known, which is precisely what the authors do. This might be a drawback because, in practice, the values of μ_1 and μ_0 are not easily available.

To derive the likelihood, consider an individual who was positive on both tests. This person could have been truly exposed and correctly tested positive on both tests ($P = \pi_1\alpha_1\alpha_2$), or the person could have been truly unexposed and falsely tested positive on both tests [$P = (1 - \pi_1)(1 - \beta_1)(1 - \beta_2\mu_0)$]. The net contribution to the likelihood is then $[\pi_1\alpha_1\alpha_2 + (1 - \pi_1)(1 - \beta_1)(1 - \beta_2\mu_0)]^{a_1}$, where a_1 is the number of individuals who tested positive on test 1 and test 2 and are cases. Similar contributions from other cells in Table 3.1 gives the

following likelihood:

$$\begin{aligned}
L &= \prod_{k=0,1} [\pi_k \alpha_1 \alpha_2 + (1 - \pi_k)(1 - \beta_1)(1 - \beta_2 \mu_0)]^{a_k} \\
&\times [\pi_k \alpha_1 (1 - \alpha_2) + (1 - \pi_k)(1 - \beta_1) \beta_2 \mu_0]^{b_k} \\
&\times [\pi_k (1 - \alpha_1)(1 - \alpha_2 \mu_1) + (1 - \pi_k) \beta_1 \beta_2]^{c_k} \\
&\times [\pi_k (1 - \alpha_1) \alpha_2 \mu_1 + (1 - \pi_k) \beta_1 (1 - \beta_2)]^{d_k}.
\end{aligned} \tag{3.1}$$

To estimate the parameters that maximize this likelihood, the authors use the EM algorithm. In this case, the missing data are the true number of exposed and the true number of unexposed subjects in each cell of Table 3.1.

A very similar methodology was developed by Hui and Walter (1980). It is a careful application of the work done on the problem of evaluating the accuracy of a new diagnostic test against a standard test with unknown error rates. This method yields a similar likelihood to (3.1). Unlike Drews et al., however, they do not introduce μ_i into their analysis and the maximization of the likelihood is done analytically. A drawback of this approach is that these estimates can sometimes lie outside the parameter space (i.e., outside the interval $[0, 1]$).

3.2.3 The Kaldor and Clayton Method

The methodology developed by Kaldor and Clayton (1985) uses latent class analysis to correct the odds-ratio estimate. To understand how their approach extends to a two-test case, we first introduce the general setup.

Suppose that all measurements made are categorical, and that the disease status D is measured without error. Suppose also that some variables V_1, \dots, V_r are measured without error, but each of the variables W_1, \dots, W_s are measurements of one of t latent classes C_1, \dots, C_t (unobservable true exposure), where $t < s$, and are subject to misclassification error. Conditional on C_k , the variables W_j are mutually independent, and independent of D and V_i . Each possible outcome in the $(1 + r + s + t)$ -dimensional table resulting from a cross-classification by all these variables can be represented by the vector $\mathbf{x} = (d, \mathbf{v}, \mathbf{w}, \mathbf{c})$, where d is disease status, and \mathbf{v}, \mathbf{w} and \mathbf{c} are respectively r -, s - and t -dimensional vectors representing categories of V_i , W_j and C_k . Further, suppose that $m_{\mathbf{x}}$ represent the number of individuals in category \mathbf{x} . Of course, this variable is unobservable, since the true values of the latent classes are unknown. One can only observe $m_{d\mathbf{v}\mathbf{w}\cdot}$, where the dot indicates summation over all categories of the t latent classes. The authors assume, however, that $m_{\mathbf{x}}$ can be generated by log-linear model of the form $\log \mu_{\mathbf{x}} = \sum_A \theta_a^A$, where $\mu_{\mathbf{x}} = E(m_{\mathbf{x}})$.

In terms of definitions above, for a model with two measurements per subject, we have $r = 0$ since there are no variables measured without error, and $t = 1$ because there is only one latent variable underlying the repeat measurements. Each of the variables D , C and W_j is dichotomous, taking the values 0 and 1. The latent class/logistic model is then of the form $\log \mu_{\mathbf{x}} = \theta + \theta_d^D + \theta_c^C + \theta_{cd}^{CD} + \sum_{j=1}^s (\theta_{w_j}^W + \theta_{cw_j}^{CW})$, where s is the number of repeat measurements made of the risk factor C . The authors utilize the EM algorithm

to estimate the relevant parameters. Here, the missing data are unobservable true exposure (C_1, \dots, C_i) .

3.3 Our Method

Our method is an extension of the Joseph, Gyorkos and Coupal (1995) method for inference on the sensitivities and specificities of two classification procedures applied to one population. We modify their method to apply to the two test, two population scenario.

Let the unobserved latent data A_i , B_i , C_i , and D_i , $i = 0, 1$ represent the number of true positive subjects out of the observed cell counts a_i , b_i , c_i , and d_i , $i = 0, 1$, respectively, in the $2 \times 2 \times 2$ table (see Table 3.1). In addition to the assumptions and definitions made in section 3.1, we will assume that the two tests determine the exposure status independently. We believe that this assumption is as reasonable as the assumption made by Drews, Flanders and Kosinski (1993), whereby they assume the degree of dependence is known.

Let us now examine the possible outcomes involving the observed and the latent data. Since any subject, whether truly exposed or not, can test positively or negatively on each test, there are eight possible combinations for both cases and controls. An individual can truly be exposed and correctly classified as exposed by both tests $[\pi_i \alpha_1 \alpha_2]$, or can truly be exposed and correctly classified by one of the tests and misclassified by the other $[\pi_i \alpha_1 (1 - \alpha_2)$ or $\pi_i (1 - \alpha_1) \alpha_2]$, and so on. Similarly, an individual can truly be unexposed and misclassified by both tests $[(1 - \pi_i)(1 - \beta_1)(1 - \beta_2)]$, or can truly be un-

exposed and correctly classified by both tests $[(1 - \pi_i)\beta_1\beta_2]$, and so forth. All eight combinations are summarized in Table 3.2.

Number of subjects	True exposure	Test 1 (T_1)	Test 2 (T_2)	Contribution to likelihood
A_i	+	+	+	$\pi_i\alpha_1\alpha_2$
B_i	+	+	-	$\pi_i\alpha_1(1 - \alpha_2)$
C_i	+	-	+	$\pi_i(1 - \alpha_1)\alpha_2$
D_i	+	-	-	$\pi_i(1 - \alpha_1)(1 - \alpha_2)$
$a_i - A_i$	-	+	+	$(1 - \pi_i)(1 - \beta_1)(1 - \beta_2)$
$b_i - B_i$	-	+	-	$(1 - \pi_i)(1 - \beta_1)\beta_2$
$c_i - C_i$	-	-	+	$(1 - \pi_i)\beta_1(1 - \beta_2)$
$d_i - D_i$	-	-	-	$(1 - \pi_i)\beta_1\beta_2$

Table 3.2: Distribution of observed and latent data when two imperfect tests are used, together with the contribution to the likelihood each combination of the observed and latent data makes. Symbol + denotes exposed, and - denotes unexposed.

Under the assumption of independence between controls and cases, the derivation of the likelihood function of the latent and observed data in Table 3.2 is straightforward.

$$\begin{aligned}
l(A_i, B_i, C_i, D_i, a_i, b_i, c_i, d_i | \pi_i, \alpha_i, \beta_i) &= \prod_{i=0,1} \frac{N_i}{A_i! \dots (d_i - D_i)!} \\
&\times [\pi_i\alpha_1\alpha_2]^{A_i} [\pi_i\alpha_1(1 - \alpha_2)]^{B_i} \\
&\times [\pi_i(1 - \alpha_1)\alpha_2]^{C_i} [\pi_i(1 - \alpha_1)(1 - \alpha_2)]^{D_i} \\
&\times [(1 - \pi_i)(1 - \beta_1)(1 - \beta_2)]^{a_i - A_i} \\
&\times [(1 - \pi_i)(1 - \beta_1)\beta_2]^{b_i - B_i} \\
&\times [(1 - \pi_i)\beta_1(1 - \beta_2)]^{c_i - C_i} \\
&\times [(1 - \pi_i)\beta_1\beta_2]^{d_i - D_i}.
\end{aligned}$$

By gathering the proper exponents and collecting the like terms, the likelihood can be rewritten as

$$\begin{aligned}
& \prod_{i=0,1} \left[\frac{N_i}{A_i! \dots (d_i - D_i)!} \pi_i^{A_i+B_i+C_i+D_i} (1 - \pi_i)^{N_i - (A_i+B_i+C_i+D_i)} \right] \\
& \times \alpha_1^{A_0+B_0+A_1+B_1} (1 - \alpha_1)^{C_0+D_0+C_1+D_1} \alpha_2^{A_0+C_0+A_1+C_1} (1 - \alpha_2)^{B_0+D_0+B_1+D_1} \\
& \times \beta_1^{c_0+d_0+c_1+d_1 - (C_0+D_0+C_1+D_1)} (1 - \beta_1)^{a_0+b_0+a_1+b_1 - (A_0+B_0+A_1+B_1)} \\
& \times \beta_2^{b_0+d_0+b_1+d_1 - (B_0+D_0+B_1+D_1)} (1 - \beta_2)^{a_0+c_0+a_1+c_1 - (A_0+C_0+A_1+C_1)}.
\end{aligned} \tag{3.2}$$

As in chapter 2, we will choose beta densities to represent the prior information available for α_i , β_i , and π_i , $i = 0, 1$. Moreover, we will assume that the six parameters are independent *a priori*, with

$$\alpha_i \sim \text{Beta}(\sigma_{\alpha_i}, \mu_{\alpha_i}), i = 1, 2,$$

$$\beta_i \sim \text{Beta}(\sigma_{\beta_i}, \mu_{\beta_i}), i = 1, 2,$$

$$\pi_0 \sim \text{Beta}(\sigma_{\pi_0}, \mu_{\pi_0}),$$

$$\pi_1 \sim \text{Beta}(\sigma_{\pi_1}, \mu_{\pi_1}),$$

giving a joint prior density

$$p(\alpha_1, \beta_1, \alpha_2, \beta_2, \pi_0, \pi_1) = \prod_{i=0,1} p_{\pi_i}(\pi_i) \prod_{j=1,2} p_{\alpha_j}(\alpha_j) p_{\beta_j}(\beta_j). \tag{3.3}$$

Since the posterior density is proportional to the likelihood function (3.2) and the prior distribution (3.3), it is easily seen that the posterior density is

proportional to

$$\begin{aligned}
& \prod_{i=0,1} \left[\frac{N_i}{A_i! \dots (d_i - D_i)!} \pi_i^{A_i+B_i+C_i+D_i+\sigma_{\pi_i}-1} (1 - \pi_i)^{N_i-(A_i+B_i+C_i+D_i)+\mu_{\pi_i}-1} \right] \\
& \times \alpha_1^{A_0+B_0+A_1+B_1+\sigma_{\alpha_1}-1} (1 - \alpha_1)^{C_0+D_0+C_1+D_1+\mu_{\alpha_1}-1} \\
& \times \alpha_2^{A_0+C_0+A_1+C_1+\sigma_{\alpha_2}-1} (1 - \alpha_2)^{B_0+D_0+B_1+D_1+\mu_{\alpha_2}-1} \\
& \times \beta_1^{c_0+d_0+c_1+d_1-(C_0+D_0+C_1+D_1)+\sigma_{\beta_1}-1} (1 - \beta_1)^{a_0+b_0+a_1+b_1-(A_0+B_0+A_1+B_1)+\mu_{\beta_1}-1} \\
& \times \beta_2^{b_0+d_0+b_1+d_1-(B_0+D_0+B_1+D_1)+\sigma_{\beta_2}-1} (1 - \beta_2)^{a_0+c_0+a_1+c_1-(A_0+C_0+A_1+C_1)+\mu_{\beta_2}-1}
\end{aligned} \tag{3.4}$$

We plan to use the Gibbs sampler to sample from (3.4). The implementation of this algorithm should be relatively straightforward as we have the following univariate conditional densities:

$$\alpha_1 | A_i, B_i, C_i, D_i, \sigma_{\alpha_1}, \mu_{\alpha_1} \sim \text{Beta}(A_0 + B_0 + A_1 + B_1 + \sigma_{\alpha_1}, C_0 + D_0 + C_1 + D_1 + \mu_{\alpha_1}),$$

$$\beta_1 | a_i, b_i, c_i, d_i, A_i, B_i, C_i, D_i, \sigma_{\beta_1}, \mu_{\beta_1} \sim \text{Beta}(c_0 + d_0 + c_1 + d_1 - (C_0 + D_0 + C_1 + D_1) + \sigma_{\beta_1}, a_0 + b_0 + a_1 + b_1 - (A_0 + B_0 + A_1 + B_1) + \mu_{\beta_1}),$$

$$\alpha_2 | A_i, B_i, C_i, D_i, \sigma_{\alpha_2}, \mu_{\alpha_2} \sim \text{Beta}(A_0 + C_0 + A_1 + C_1 + \sigma_{\alpha_2}, B_0 + D_0 + B_1 + D_1 + \mu_{\alpha_2}),$$

$$\beta_2 | a_i, b_i, c_i, d_i, A_i, B_i, C_i, D_i, \sigma_{\beta_2}, \mu_{\beta_2} \sim \text{Beta}(b_0 + d_0 + b_1 + d_1 - (B_0 + D_0 + B_1 + D_1) + \sigma_{\beta_2}, a_0 + c_0 + a_1 + c_1 - (A_0 + C_0 + A_1 + C_1) + \mu_{\beta_2}),$$

$$\pi_i | A_i, B_i, C_i, D_i, a_i, b_i, c_i, d_i, \sigma_{\pi_i}, \mu_{\pi_i} \sim \text{Beta}(A_i + B_i + C_i + D_i + \sigma_{\pi_i}, N_i - (A_i + B_i + C_i + D_i) + \mu_{\pi_i}), i = 0, 1,$$

$$A_i | \pi_i, \alpha_i, \beta_i, a_i \sim \text{Binomial}(a_i, \frac{\pi_i \alpha_1 \alpha_2}{\pi_i \alpha_1 \alpha_2 + (1 - \pi_i)(1 - \beta_1)(1 - \beta_2)}), i = 0, 1,$$

$$B_i | \pi_i, \alpha_i, \beta_i, b_i \sim \text{Binomial}(b_i, \frac{\pi_i \alpha_1 (1 - \alpha_2)}{\pi_i \alpha_1 (1 - \alpha_2) + (1 - \pi_i)(1 - \beta_1) \beta_2}), i = 0, 1,$$

$$C_i | \pi_i, \alpha_i, \beta_i, c_i \sim \text{Binomial}(c_i, \frac{\pi_i (1 - \alpha_1) \alpha_2}{\pi_i (1 - \alpha_1) \alpha_2 + (1 - \pi_i) \beta_1 (1 - \beta_2)}), i = 0, 1,$$

$$D_i | \pi_i, \alpha_i, \beta_i, d_i \sim \text{Binomial}(d_i, \frac{\pi_i (1 - \alpha_1)(1 - \alpha_2)}{\pi_i (1 - \alpha_1)(1 - \alpha_2) + (1 - \pi_i) \beta_1 \beta_2}), i = 0, 1,$$

We see that the Gibbs sampler enables us to draw inference not only on the prevalences of exposure (and consequently on the odds-ratio), but also on the the sensitivities and specificities of the two classification schemes.

3.3.1 Examples

Here, we plan to investigate the validity of our method by applying it to eight hypothetical case-control studies where the true values of sensitivities, specificities, prevalence of exposure in controls and prevalence of exposure in cases are known. Data will be simulated using these true values for three different sample sizes, $N_0 = N_1 = 200$, $N_0 = N_1 = 800$ and $N_0 = N_1 = 3200$. One example will be discussed in more detail and the the results of the remaining eight will be summarized in a table.

Example 1 Consider a scenario where the true values of the six parameters are $(\alpha_1, \beta_1, \alpha_2, \beta_2, \pi_0, \pi_1) = (0.85, 0.90, 0.90, 0.88, 0.07, 0.18)$. The value of π_0 and π_1 imply $\phi = 2.92$ or $\log \phi = 1.07$. Data was simulated under this scenario for three different sample sizes. For the sample size $N_0 = N_1 = 200$, simulated values were $(a_0, b_0, c_0, d_0) = (11, 9, 21, 159)$ and $(a_1, b_1, c_1, d_1) = (33, 8, 13, 146)$. For the sample size $N_0 = N_1 = 800$, simulated values were

$(a_0, b_0, c_0, d_0) = (47, 83, 90, 580)$ and $(a_1, b_1, c_1, d_1) = (113, 78, 90, 519)$. Finally, for the sample size $N_0 = N_1 = 3200$, simulated values were $(a_0, b_0, c_0, d_0) = (181, 295, 348, 2376)$ and $(a_1, b_1, c_1, d_1) = (453, 297, 338, 2112)$. As in chapter 2, the prior for the prevalence of exposure in controls and the prevalence of exposure in cases was chosen to be $Beta(1, 1)$. Furthermore, suppose that the investigator does not have a very good prior knowledge about the sensitivities and specificities of the two test, but is only willing to assume that the probabilities of correct classification are greater than chance (i.e., they are in the interval $[0.5, 1]$). We will translate this information into a Beta prior which assigns a 0.95 probability to this interval and with its mode at a value in this interval. In this example, we will set the mode at 0.8 for all four classification probabilities. This implies a $Beta(7.55, 2.64)$ prior for the sensitivities and specificities. Five independent chains of the Gibbs sampler were run using different starting values. The sequential output is shown in Figure 3.1.

Before attempting any inference, we will first examine the adequacy of the Gibbs sampler. The sequential plots of the post burn-in period (the first 500 observations from each chain were discarded) in Figure 3.1 indicate that convergence has occurred. The five independent chains appear to consistently converge to the same region in the parameter space, with no instances of slow mixing or drifting behaviour. Hence, we will regard Figure 3.1 as informal evidence that our posterior sampling is adequate. The statistical inference will be based on $5 \times (2500 - 500) = 10000$ sampled pairs of the prevalence of exposure in controls and prevalence of exposure in cases.

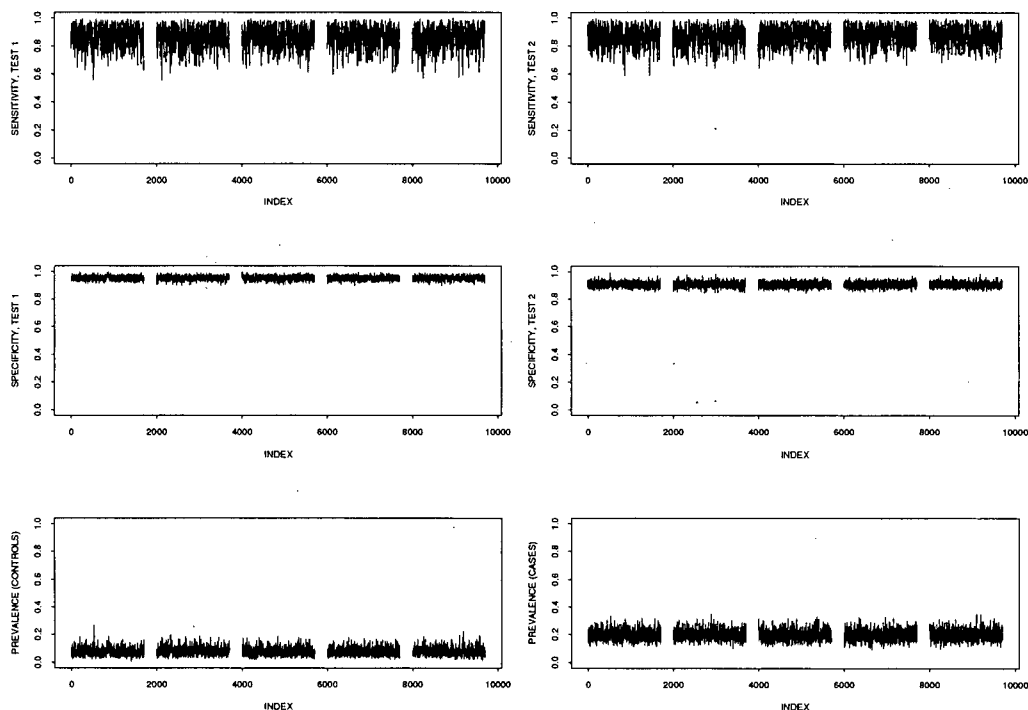


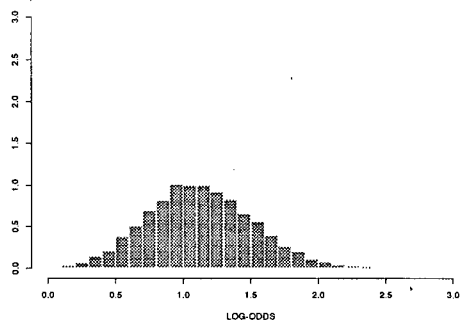
Figure 3.1: Post burn-in output of the five independent chains of the Gibbs sampler for sensitivities, specificities, prevalence of exposure in controls and prevalence of exposure in cases in Example 1. Data was simulated for $N_0 = N_1 = 200$.

Figure 3.2 shows the histograms of the posterior log odds-ratio for the three sample sizes considered, together with the true log odds-ratio, which is obtained when $N_i \rightarrow \infty$, $i = 0, 1$. We see that the convergence toward the true value of log odds-ratio is much more rapid than in the case when one test is used. Also, as there is no nonidentifiability in the likelihood, the estimation error is $O(n^{-1/2})$, which is witnessed in the widths of the posterior distributions. As the sample size increases from 200 to 800, and from 800 to

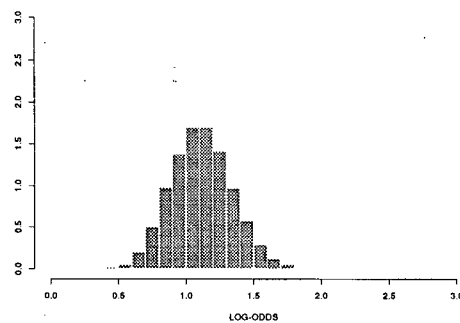
3200, the width of the posterior distribution decreases roughly by the factor of 2.

The posteriors in Figure 3.2 are, of course, based on single datasets. To assess the generality of the findings and the coverage of the credible intervals, we simulate 1000 datasets under the same scenario for the three sample sizes. For each dataset and each sample size we compute 80% HPD credible intervals for $\log \phi$. The first forty intervals appear in Figure 3.3. The average interval widths are 1.01, 0.51, and 0.28 and the empirical coverage rates for the intervals are 85.7%, 83.8%, and 84.1% for the sample sizes $N_0 = N_1 = 200$, $N_0 = N_1 = 800$, and $N_0 = N_1 = 3200$, respectively. As the standard error of the empirical coverage rates is approximately 1.26, it appears that 80% HPD credible intervals exhibit slight overcoverage for all three sample sizes.

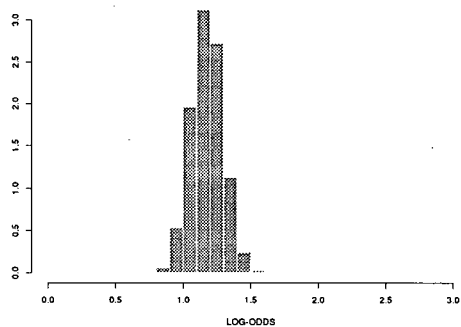
When doing Bayesian inference, it is important to assess the sensitivity of estimates. As we have already seen in Figure 3.1, the Gibbs sampler is not at all sensitive to the choice of starting values; all five chains converged to the same region in the parameter space. We now investigate whether the estimates of $\log \phi$ are overly sensitive to different choices of the hyperparameters. We re-run the Gibbs sampler algorithm for the intermediate sample size, $N_0 = N_1 = 800$, with different values of the hyperparameters. We consider two diametrically opposed scenarios. First, we examine the case where $Beta(1, 1)$ prior is used for all six parameters. This scenario could potentially arise when an investigator has no prior knowledge of any of the six parameters, or is not certain of the validity of the available information. Second, we consider the



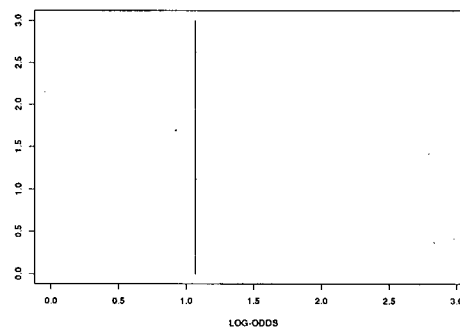
(a) Sample size: $N_0 = N_1 = 200$



(b) Sample size: $N_0 = N_1 = 800$



(c) Sample size: $N_0 = N_1 = 3200$



(d) Sample size: $N_0 = N_1 = \infty$

Figure 3.2: Posterior samples of the logarithm of the odds-ratio for the datasets in Example 1.

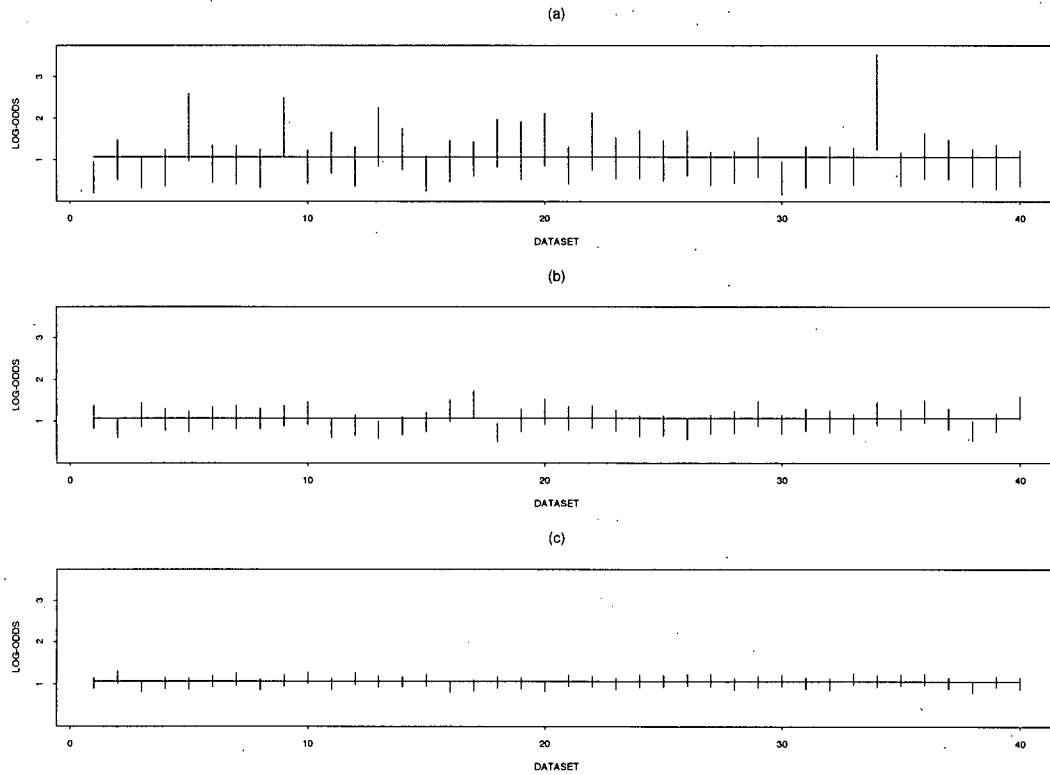


Figure 3.3: 80% highest posterior density credible intervals for $\log \phi$ in Example 1. The solid vertical lines represent credible intervals for forty data sets and the solid horizontal line indicates the true value of $\log \phi$. The panel (a) gives the intervals for the sample size $N_0 = N_1 = 200$, the panel (b) gives the intervals for the sample size $N_0 = N_1 = 800$, and the panel (c) gives the intervals for the sample size $N_0 = N_1 = 3200$.

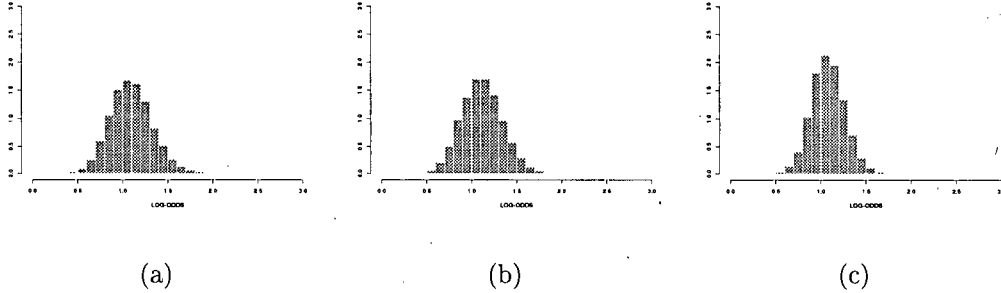


Figure 3.4: Posterior samples of the logarithm of the odds-ratio for the sample size $N_0 = N_1 = 800$ in Example 1. (a) $Beta(1, 1)$ prior used for all six parameters. (b) $Beta(7.55, 2.64)$ prior used for α_i and $\beta_i, i = 1, 2$, while $Beta(1, 1)$ prior used for π_0 and π_1 (the histogram is identical to that shown in Figure 3.2 (b)). (c) $Beta$ prior for each parameter was chosen so that the mode γ is centred at the true value and 0.95 probability is assigned to the interval $\gamma \pm 0.05$.

case where very good prior knowledge is available for all six parameters. The hyperparameters σ and μ of $Beta(\sigma, \mu)$ priors were chosen so that the mode γ is centred at the true value of each parameter and 0.95 probability is assigned to the interval $\gamma \pm 0.05$. The histograms of $\log \phi$ are shown in Figure 3.4, together with the histogram identical to that shown in Figure 3.2 (b).

As evidenced in Figure 3.4, the posterior distribution of $\log \phi$ is very insensitive to the choice of hyperparameter values in this example, suggesting that the precise knowledge of hyperparameter values is not very important to the analysis.

The results for the remaining eight examples are shown in Table 3.3. Under each scenario, 1000 data sets were simulated to assess the coverage of

the 80% HPD credible intervals and the mean interval length. It appears that for each scenario and each sample size, 80% HPD credible intervals exhibit slight overcoverage.

Parameters					Empirical coverage of 80% HPD CI		
α_1	β_1	α_2	β_2	$\log \phi$	$N_i = 200$	$N_i = 800$	$N_i = 3200$
0.96	0.96	0.96	0.96	0.75	83.3% (0.94)	82.8% (0.48)	81.9% (0.25)
0.95	0.90	0.90	0.95	0.75	83.2% (0.96)	84.3% (0.47)	83.1% (0.27)
0.90	0.90	0.90	0.87	0.86	84.4% (0.98)	84.7% (0.50)	84.2% (0.28)
0.85	0.85	0.87	0.87	0.86	83.8% (1.03)	83.6% (0.51)	82.1% (0.24)
0.85	0.83	0.87	0.85	1.54	81.9% (1.02)	84.9% (0.52)	83.0% (0.22)
0.82	0.82	0.82	0.82	1.54	81.8% (1.03)	82.6% (0.52)	82.5% (0.29)
0.85	0.90	0.90	0.88	1.07	83.2% (1.01)	82.5% (0.47)	82.6% (0.28)
0.65	0.65	0.65	0.65	1.11	81.2% (1.04)	82.8% (0.49)	81.6% (0.28)

Table 3.3: True parameter values and the empirical coverage of the 80% HPD credible intervals for $\log \phi$ for the sample sizes $N_i = 200$, $N_i = 800$, and $N_i = 3200$, $i = 0, 1$. Numbers in brackets represent mean interval length. For the empirical coverage and the mean length, 1000 data sets were simulated.

From the above table, we see that the empirical coverage rates do not appear to depend on the values of the six parameters. In particular, the empirical coverage rates for the scenario where relatively high values were chosen for the sensitivities and specificities (the first row in Table 3.3) are not significantly different than the empirical coverage rates for the scenario where relatively low values were chosen for the sensitivities and specificities (the last row in Table 3.3).

3.4 An Application to Real Data

We will now compare our method to the method of Drews et al. (1993) by analyzing data from a case-control study of sudden infant death syndrome (SIDS). The exposure data was obtained from maternal interviews and medical records. The data appear in Drews et al. (1993), as taken from Hoffman et al. (1988). These data come from the National Institute of Child Health and Human Development's (NICHD) case-control study of SIDS. A total of 844 SIDS victims and two age-matched living infants were included in the case-control study. A first control was individually age-matched to the case in such a way that, at the time of the interview, the control would be the same age as the case had been when he/she died. A second control was matched to the case on birth-weight and race. Drews et al. (1993) were given access to a data on a pseudorandom sample (a systematic sample with a random start) of 226 of the 844 SIDS cases and one of the two age-matched controls for each of the 226 cases. See Hoffman et al. (1988) or Drews, Kraus and Greenland (1990) for a full description.

Mothers of cases and controls were interviewed regarding events which had occurred during their pregnancies, their labour and deliveries, and their infants lives within five weeks of the death of the cases.

We looked at six dichotomous exposure variables: maternal anemia during pregnancy (ANEM), maternal urinary tract infection during pregnancy (UTI), previous spontaneous abortion (PSA), low pregnancy weight gain, which was defined as weight gain of less than 15 pounds (LPWG), mater-

		Controls		Cases	
		MR		MR	
		+	-	+	-
ANEM					
	Int +	20	43	24	49
	Int -	15	147	15	125
UTI					
	Int +	14	10	13	14
	Int -	4	190	14	174
PSA					
	Int +	21	13	23	12
	Int -	11	175	15	169
PV					
	Int +	91	12	69	9
	Int -	6	66	9	78
LPWG					
	Int +	6	3	11	13
	Int -	4	86	5	73
PAU					
	Int +	21	12	29	22
	Int -	16	168	17	143

Table 3.4: A pseudorandom sample of 226 SIDS cases and 226 controls from the NICHD study. Data was classified using medical record (MR) and interview (Int) data.

nal antibiotic use during pregnancy (PAU), and polio vaccination (PV) before death or interview. Data for these variables were available from both medical records and maternal interviews. The subjects for which information is missing from either source were not included in the analysis. Further, since the matching of cases and controls was fairly weak, it was ignored. Table 3.4 shows the distribution of controls and cases according to both interview and medical record data for the six variables. Interview data were arbitrarily taken to represent test 1 and medical records to represent test 2. This distinction has no ill effect, since we assumed that the errors between two tests are independent.

As no reliable prior information about the sensitivity and specificity of the interview data and medical records was available, we assigned a $Beta(1, 1)$ prior to all six parameters. The results of our analysis are shown in Table 3.4 and Table 3.5. Table 3.4 shows median estimates of the sensitivities, specificities and $\log \phi$, together with 95% HPD credible intervals for $\log \phi$. Table 3.5 shows 95% HPD credible intervals for the sensitivities and specificities of the interview data and medical records. Five independent chains of the Gibbs sampler were run with different starting values. The inference was based on $5 \times (5500 - 500) = 25000$ draws. As mentioned before, Drews et al. (1993) used the EM algorithm to obtain estimates of the sensitivity and specificity of the interview data and medical records and the odds-ratio. The results of their analysis appear in Table 3.6.

Note that for data set, at least one of the parameters estimated using Drews et al. (1993) method lies on the boundary of the parameter space (i.e.,

1.00). This is, of course not true when the Gibbs sampler is used. Except for this distinction, the EM algorithm and the Gibbs sampler produce similar estimates of the sensitivities and specificities for the ANEM, PV, LPWG, and PAU data sets. Greater discrepancies occur for the UTI and PSA datasets. For instance, in the UTI dataset, the estimate of the sensitivity of medical records obtained by the Gibbs sampler is 0.67, whereas the estimate obtained using the EM algorithm is 0.56, a difference of 0.11. Also, the estimate of the sensitivity of interview data using the Gibbs sampler is 0.70, while the one obtained via the EM algorithm is 0.60. A similar finding applies to the PSA dataset.

The point estimates of $\log \phi$ are similar in magnitude, though the ones obtained via the Gibbs sampler are smaller than the EM estimates for all but the ANEM data set. However, the credible intervals are narrower than the confidence intervals for all six data sets. In particular, in five out of six data sets, the credible intervals for $\log \phi$ are completely contained in the confidence intervals. This finding is somewhat surprising, since our credible intervals take into account the variability associated with the sensitivities and specificities of the two tests, which is not true for the confidence intervals. The variance used to calculate the confidence interval for $\log \phi$ is estimated using the delta method, as follows:

$$\text{Var}(\log \phi) = \frac{\text{Var}(\pi_0)}{[\pi_0(1 - \pi_0)]^2} + \frac{\text{Var}(\pi_1)}{[\pi_1(1 - \pi_1)]^2} - \frac{2\text{Cov}(\pi_0, \pi_1)}{[\pi_0(1 - \pi_0)][\pi_1(1 - \pi_1)]}.$$

The values of $\text{Var}(\pi_0)$, $\text{Var}(\pi_1)$, and $\text{Cov}(\pi_0, \pi_1)$ can be obtained by substituting the parameter estimates into the expected information matrix and in-

verting this matrix to obtain the estimated covariance matrix. Therefore, $\text{Var}(\log \phi)$ only depends on the point estimates of the sensitivities and specificities, not the estimates of their standard errors. The explicit formulae for the information matrix when the errors of the two classification schemes are independent ($\mu_1 = \mu_0 = 1$) are given in Hui and Walter (1980).

Since we have established the coverage of the credible intervals through the simulation studies, and the length via the analysis of real data set, we can conclude that they perform better than the confidence intervals.

3.5 Discussion

In this chapter we have examined the situation where the exposure status is determined by two imperfect classification schemes. We have seen that the Gibbs sampler works very well in this case, presumably because the likelihood is identifiable. Our method performed well both in the simulation studies and in the analysis of the real data, where it appeared to outperform the EM algorithm approach of Drews et al. (1993).

Variable	Interview Data		Medical Record		$\log \phi$	95% HPD Cr. Int for $\log \phi$
	α	β	α	β		
ANEM	0.75	0.91	0.42	0.94	0.52	(-0.07, 1.15)
UTI	0.70	0.97	0.67	0.98	0.45	(-0.39, 1.07)
PSA	0.75	0.96	0.77	0.96	0.15	(-0.42, 0.74)
PV	0.96	0.94	0.92	0.94	-0.43	(-0.89, 0.01)
LPWG	0.78	0.98	0.56	0.97	1.08	(0.15, 2.01)
PAU	0.77	0.96	0.67	0.94	0.58	(0.02, 1.17)

Table 3.5: Median estimates of the sensitivities, specificities and the logarithm of the odds-ratio, and 95% HPD credible intervals for the logarithm of the odds-ratio, using the Gibbs sampler.

Variable	Interview Data		Medical Record	
	α	β	α	β
ANEM	(0.61, 0.90)	(0.83, 0.96)	(0.30, 0.54)	(0.88, 0.97)
UTI	(0.56, 0.88)	(0.91, 1.00)	(0.49, 0.87)	(0.91, 1.00)
PSA	(0.63, 0.89)	(0.89, 0.99)	(0.64, 0.92)	(0.90, 0.99)
PV	(0.89, 0.99)	(0.87, 0.98)	(0.84, 0.97)	(0.88, 0.98)
LPWG	(0.60, 0.91)	(0.93, 1.00)	(0.41, 0.74)	(0.93, 1.00)
PAU	(0.61, 0.91)	(0.89, 1.00)	(0.49, 0.87)	(0.89, 0.98)

Table 3.6: 95% HPD credible intervals for the sensitivities and specificities of the interview data and the medical records.

Variable	Interview Data		Medical Record		$\log \phi$	95% Conf. Int for $\log \phi$
	α	β	α	β		
ANEM	0.78	1.00	0.35	0.93	0.51	(-0.23, 1.24)
UTI	0.60	0.99	0.56	1.00	0.51	(-0.27, 1.30)
PSA	0.63	0.93	1.00	1.00	0.21	(-0.44, 0.86)
PV	1.00	1.00	0.88	0.91	-0.46	(-0.94, 0.02)
LPWG	0.84	1.00	0.52	0.97	1.17	(0.2, 2.31)
PAU	0.79	1.00	0.60	0.94	0.62	(-0.01, 1.24)

Table 3.7: Estimates of the sensitivities, specificities and the logarithm of the odds-ratio, and 95% confidence intervals for the logarithm of the odds-ratio, using the EM algorithm (Drews et al. (1993)).

Chapter 4

Conclusion

As we have seen, misclassification of exposure poses serious problems in statistical analysis. The effects of ignoring misclassification and the methods which correct for it have received a considerable attention in literature. More often than not, exposure misclassification substantially biases estimates of the relative risk, even when misclassification rates are very small. Nondifferential misclassification, for example, tends to attenuate observed exposure-disease relationships (odds-ratio, for instance).

Corrected estimates of the odds-ratio ϕ can be easily obtained if the misclassification probabilities (the sensitivity and specificity) are known. Baron (1977) and Greenland and Kleinbaum (1983) provide classical methods for correcting the odds-ratio, which rely on the invertibility of matrices. We introduced a Bayesian approach to correcting the odds-ratio when the sensitivity and specificity are known. The sampling from the posterior distribution of prevalence of exposure among cases and control (and hence, the odds-ratio)

proved to be quite simple.

However, having the exact values of the sensitivity and specificity is not very common in practice. Instead, good guesses of these values might be available (through previous studies or by comparing the diagnostic test to a gold standard). It might be very tempting to carry out the above analysis by pretending that these guesses are the true values of the sensitivity and specificity. We showed, however, that even the smallest difference between the true and guessed values can result in very large difference between the miscorrected and true odds-ratio.

This result, together with the previous work of Marshall (1989), suggested that it is reasonable to incorporate the uncertainty into the analysis. Bayesian methods make this feasible, as partial knowledge about the sensitivity and specificity is easily represented with the appropriate prior distribution. We suggested two algorithms to sample from the posterior distribution: the Gibbs sampler and the Metropolis-Hastings. The Gibbs sampler did not work particularly well. The possible explanation is that drifting behaviour of the Gibbs sampler may occur when it is applied to the nonidentifiable model. On the other hand, the Metropolis-Hastings algorithm applied to a parameterization that separates the identifiable and nonidentifiable part seemed to perform well. We were able to demonstrate that the approach which admits the uncertainty in the guesses of the sensitivity and specificity performs better than the method that ignores it. We also showed that, despite nonidentifiability, the marginal prior and posterior distributions of (α, β) were not equal. As a

result of learning about the apparent exposure prevalences (θ_1, θ_2) , one could learn about (α, β) .

In chapter 3 we examined the scenario when the exposure data is available from two imperfect sources. Various methods for improving the odds-ratio estimates by combining data from two imperfect classification schemes have been suggested, of which we outlined three. The method of Marshall and Graham (1984) uses only the concordant observations to correct the odds-ratio estimates. The methods of Drews et al. (1993) and Kaldor and Clayton (1985) are both classical approaches that incorporate the underlying true exposure into their analysis. We developed a Bayesian latent class approach, where the latent class was the true exposure. We used the Gibbs sampler to sample from the posterior distributions of the six parameters. Both the adequacy of the Gibbs sampler and the validity of our method were established through a simulation study.

The analysis of case-control data on the sudden infant death syndrome showed that our method could easily be applied in practice. The analysis also showed that our method appears to outperform the method of Drews et al. (1993).

The following suggestions could be considered for further development of the methodology presented here. One would be to extend the method to the situation where three or more imperfect tests are used to assess the exposure. This should be a very simple extension of our methodology. Other would be to modify the methods to apply to matched-pair case-control studies. And finally,

it seems useful to explore the situation where, in addition to the misclassified dichotomous exposure, other covariates are measured, possibly without error. This does not appear to be a matter of simple extension of our method, because it would require the use of a link function between the probability of disease and covariates in question.

Bibliography

- [1] Barron, B.A. "The effects of misclassification on the estimation of relative risk." *Biometrics*, **33**, 414-418 (1977).
- [2] Blettner, M., Wahrendorf, J. "What does an observed relative risk convey about possible misclassification?." *Methods of Information in Medicine*, **41**, 923-937 (1984).
- [3] Chib, S., Greenberg, E. "Understanding the Metropolis-Hastings algorithm." *The American Statistician*, **49**, 327-335 (1995).
- [4] Dawid, A.P. "Conditional independence in statistical theory (with discussion)." *Journal of the Royal Statistical Society B*, **41**, 1-31 (1979).
- [5] Dempster, A.P., Laird, N.M., Rubin, D.B. "Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society B*, **39**, 1-38 (1977).
- [6] Drews, C.D., Flanders, W.D., Kosinski, A.S. "Use of two data sources to estimate odds ratios in case-control studies." *Epidemiology*, **4**, 327-335 (1993).

- [7] Drews, C.D., Kraus, J.R., Greenland, S. "Recall bias in a case-control study of sudden infant death syndrome." *International Journal of Epidemiology*, **19**, 405-411 (1990).
- [8] Gelfand, A.E., Sahu, S.K. "Identifiability, improper priors, and Gibbs sampling for generalized linear models." *Journal of the American Statistical Association*, **94**, 247-253 (1999).
- [9] Gentle, J.E. *Random Number Generation and Monte Carlo Methods*. New York, Springer-Verlag, 1998.
- [10] Greenland, S., Kleinbaum, D.G. "Correcting for misclassification in two-way tables and matched-pair studies." *International Journal of Epidemiology*, **12**, 93-97 (1983).
- [11] Hoffman, H.J., Hunter, J.C., Ellish, N.J., Janerich, D.T., Goldberg, J. "Adverse reproductive factors and the sudden infant death syndrome." In: Harper, C.M.R., Hoffman, H.J., eds. *Sudden Infant Death Syndrome: Risk Factors and Basic Mechanisms*. New York, PMA Publishing, 1988.
- [12] Hui, S.L., Walter, S.D. "Estimating the error rates of diagnostic tests." *Biometrics*, **36**, 167-171 (1980).
- [13] Joseph, L., Gyorkos, T.W., Coupal, L. "Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard." *American Journal of Epidemiology*, **141**, 263-272 (1995).

- [14] Kaldor, J., Clayton, D. "Latent class analysis in chronic disease epidemiology." *Statistics in Medicine*, **4**, 327-335 (1985).
- [15] Marshall, J.R. "The use of dual or multiple reports in epidemiological studies." *Statistics in Medicine*, **8**, 1041-1049 (1989).
- [16] Marshall, J.R., Graham, S. "Use of dual responses to increase validity of case-control studies." *Journal of Chronic Diseases*, **37**, 125-136 (1984).
- [17] Thomas, D., Stram, D., Dwyer, J. "Exposure measurement error: influence on exposure-disease relationships and methods of correction." *Annual Review of Public Health*, **14**, 69-93 (1993).
- [18] Walter, S.D. "Commentary on 'Use of dual responses to increase validity of case-control studies' ." *Journal of Chronic Diseases*, **37**, 137-139 (1984).
- [19] Walter, S.D., Irwig, L.M. "Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review." *Journal of Clinical Epidemiology*, **9**, 923-937 (1988).