ABSTRACT

Bayesian Approaches to Parameter Estimation and Variable Selection for Misclassified Binary Data

Daniel Beavers, Ph.D.

Chairperson: James D. Stamey, Ph.D.

Binary misclassification is a common occurrence in statistical studies that, when ignored, induces bias in parameter estimates. The development of statistical methods to adjust for misclassification is necessary to allow for consistent estimation of parameters. In this work we develop a Bayesian framework for adjusting statistical models when fallible data collection methods produce misclassification of binary observations. In Chapter 2, we develop an approach for Bayesian variable selection for logistic regression models in which there exists a misclassified binary covariate. In this case, we require a subsample of gold standard validation data to estimate the sensitivity and specificity of the fallible classifier. In Chapter 3, we propose a Bayesian approach for the estimation of population prevalence of a biomarker in repeated diagnostic testing studies. In such situations, it is necessary to account for interindividual variability which we achieve through both the inclusion of random effects within logistic regression models and Bayesian hierarchical modeling. Our examples focus on applications for both reliability studies and biostatistical studies. Finally, we develop an approach to attempt to detect conditional dependence parameters between two fallible diagnostic tests for a binary logistic regression covariate in the absence of a gold standard test in Chapter 4. We compare the performance of the proposed procedure to previously published means assessing model fit.

Bayesian Approaches to Parameter Estimation and Variable Selection for Misclassified Binary Data

by

Daniel Beavers, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

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

Approved by the Dissertation Committee

James D. Stamey, Ph.D., Chairperson

Thomas L. Bratcher, Ph.D.

Rafer S. Lutz, Ph.D.

John W. Seaman Jr., Ph.D.

Dean M. Young, Ph.D.

Accepted by the Graduate School
August 2009

J. Larry Lyon, Ph.D., Dean

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you Dr. Stamey for working with me throughout this process and for your endless enthusiasm. You were a fantastic mentor, and I greatly appreciate all the time you sacrificed listening to me talk about anything and everything. Thank you Dr. Bratcher for convincing me that I should come to Baylor to get a Ph.D. No academic program has made me feel as welcome from the very beginning as the Baylor statistics department, and you were an enormous part of that. Thank you to Dr. Seaman for being an example of the scientist I hope to someday become and to Dr. Tubbs for always being incredibly helpful with my unending questions. In general, thanks to the entire department faculty for being so unselfishly committed to instilling a knowledge and excitement for statistics that I doubt I ever would have known on my own. Thank you to Dr. Young, Phil, Johnny, Brandi, and all others who were always ready and willing to talk with me about anything but statistics. What will I ever do without you? Thanks Mom and Dad for all your love and support over the years. Finally, thank you to Kristen for undertaking this journey with me and especially for your somehow contagious work ethic. I'm not sure I ever would have even pursued this, much less finished this, without you.

DEDICATION

To Kristen,

My best friend,

My marathon training partner, and

My beloved wif

CHAPTER ONE

Introduction

1.1 Misclassification

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

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

There are multiple ways to model the relationship between T and X, including

- When a gold standard classifier for T exists but its use is excessively expensive or may introduce unnecessary risks, it may be desirable to use a double-sampling protocol. In this case, all individuals are classified using one fallible test protocol, summarized by X, and a smaller subsample of these individuals are classified a second time using the gold standard test T. Using the relationship of X and T from the subsample, we can determine the misclassification rates of X and adjust the complete data set for individuals from the larger sample who have X but no T. This approach will be further explored in Chapter 2.

- When T is unobservable, a multiple test protocol may be utilized in which vector X contains multiple fallible classifications of an individual. One of the most common and efficient means is the dual test protocol, in which X_1 is measured with sensitivity and specificity S_1 and C_1, respectively, and X_2 likewise has respective sensitivity and specificity S_2 and C_2. The statistical issues regarding this form of adjustment for identifiability primarily concern parameter identifiability and possible conditional dependence between the two diagnostic tests. These issues are discussed in Chapter 4.

- When T is unobservable, repeated independent binary tests may be performed on an observation. Each test has the same sensitivity and specificity, and, provided that at least a moderately large number of individuals have at minimum three repeated tests, we can produce estimates of the sensitivity, specificity, and prevalence τ. This is the approach taken in Chapter 3.

1.2 Bayesian Estimation

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

The Bayesian approach to statistical inference differs from the frequentist approach in the sense that θ is assumed to be a random variable rather than a fixed quantity. As a random variable, we define θ to have range Θ as well as its own density function $p(\theta)$. We refer to this function as the *prior* distribution which like any other density function contains all known information about θ. Inference from a Bayesian perspective thus combines our prior information of the parameter with the observed knowledge gained from the likelihood using Bayes' rule to yield a posterior distribution

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

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

The limitations to Bayesian inference historically dealt primarily with the often complicated and intractable form of the posterior distribution $p(\theta|\mathbf{y})$. However, the advent of Markov Chain Monte Carlo (MCMC) methods have allowed practitioners of Bayesian inference to obtain numerical representations of the posterior distribution via computationally intensive tools such as the Gibbs Sampler and Metropolis-Hastings algorithm. A thorough treatment of MCMC methods can be found in Robert and Casella (2004).

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

1.3 Modeling Probabilities Associated with Binary Outcomes

Modeling data with binary outcomes typically utilizes three link functions to relate the response probability $\pi_i = pr(Y_i = 1)$ to the covariate vector $\mathbf{z}_i$:

(1) the logit link:

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

(2) the probit link:

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

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

(3) the log link:

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

We will proceed using the logit link function throughout this document due to its ease of interpretation of model parameters. The estimated value of parameter β_k associated with covariate z_k is interpreted as the change in the log-odds of the response per unit change in z_k, where we define the odds as $\pi/(1-\pi)$. The interpretation is not quite as straightforward using alternative link functions, and the use of the logit link is nearly ubiquitous in the public health sciences. More information on modeling discrete outcomes and the use of logistic regression can be found in Agresti (2002) Chapters 4 and 5.

1.4 Model Selection from a Bayesian Perspective

In Chapters 2 and 4 we explore the process of variable selection from a Bayesian perspective. A worthwhile introduction and summary of current methods is found in Dellaportas et al. (2002), which provides a suitable background for the methods discussed in this dissertation. Model selection has been a prominent if not often poorly defined branch of statistical modeling that attempts to identify the single model out of a larger set of possible models that satisfies certain optimal properties. We refer to the process as often poorly defined because of the broad diversity of optimal properties as well as the subjective definition of the possible models under consideration.

We briefly introduce variable selection from a Bayesian perspective in a similar manner to Dellaportas et al. (2002). Suppose that we observe data Y that we assume was generated by one of M possible models. We represent the (typically finite) set of possible arrangement of parameters as Θ_M such that $\theta_m \in \Theta_M$ where $m = 1, \ldots, M$. It is our desire to identify the posterior probability of the model mth model $p(m|y)$ based on the prior probability of that model $p(m)$ and the likelihood of the model given the data $p(y|m)$ using Bayes theorem

$$p(m|y) = \frac{p(m)p(y|m)}{\sum_{i=1}^{M} p(m)p(y|m)}. \tag{1.1}$$

Once we have identified $p(m|y)$ for all $m \in M$ we can identify the model that best meets our optimal properties by simply identifying

$$\max_{m \in M}\{p(m|y)\} = m'.$$

Assuming model m' has parameters $\theta_{m'}$, we can then uniquely estimate $\theta_{m'}$.

We note based on the denominator of 1.1 that the posterior probability of the mth model depends on the other possible models being considered. This leaves open the possibility that $\max_{m \in M}\{p(m|y)\} = m'$ as a consequence of the elements of Θ rather than as a consequence of model m' being the data generating model. Thus, it is important that if we are attempting to recover the data generating model that it is one of the models under consideration by our approach, which potentially increases the possible size of Θ_M. The size of Θ_M is a difficult matter to manage. If we are considering including k possible covariates for consideration in the model, then there are 2^k possible unique arrangements of parameters. Our proposed method in Chapter 2 attempts to identify the highest posterior probability model in the presence of a moderately large k.

The actual size of M is often limited by the number of covariates observed/available and practical concerns such as whether to consider all possible arrangements of interactions of covariates and polynomial expansions. In the broader context of model selection, the larger question exists whether the basic parametric assumptions are suitable and whether linear or nonlinear fits are more appropriate. Furthermore, even if we include the "data generating" model among the set Θ, the ability to detect a covariate as significantly contributing to the posterior is often a function of the sample size, regardless of the selection criteria utilized.

1.5 Identifiability

We define model *identifiability* using Casella and Berger (2002) such that for a parameter θ and θ' from the same family of distributions with density/mass function $f(\cdot)$, if $\theta \neq \theta'$, then $f(x|\theta) \neq f(x|\theta')$. Identifiability becomes a concern typically when the number of parameters exceed the available data for their estimation. In the typical assumption for analysis of variance for j groups, we express the mean of the ith group θ_i as

$$\theta_i = \mu + \tau_i,$$

where μ is the common mean of all groups and τ_i is the deviation from the common mean for the ith group. We cannot uniquely estimate μ and τ_i because we have j groups and $j+1$ parameters, requiring the simplifying assumption that $\sum_{i=1}^{j} \tau_i = 0$.

With regard to misclassification, in section 1.1 we mention that we cannot correct for the effect of misclassification when we only observe X, our fallible classifier. This occurs because we define the relationship of the true classification status T to

X as

$$pr(X = 1|T) = T \times S + (1 - T) \times (1 - C),$$

where S is our sensitivity and C is the specificity. If we know T we have observed two random variables to estimate the two unknown probabilities S and C. Without knowledge of T this further reduces to

$$pr(X = 1) = E_T[pr(X = 1|T)] = \tau \times S + (1 - \tau) \times (1 - C),$$

which involves three unknown parameters.

Identifiability is as much of a concern to a Bayesian statistician as a frequentist, although the Bayesian approach has the unique ability to estimate models that lack identifiability under certain conditions. Specifically, if we have information about a parameter that we can represent in the prior distribution, we can produce posterior estimates of model parameters that combine these known and unknown portions. We confront further identifiability issues in Chapter 4.

The chapters are divided as follows: in Chapter 2 we introduce a Bayesian variable selection in the presence of a misclassified binary covariate. In Chapter 3 we discuss a Bayesian approach to parameter estimation using pass/fail data in which we introduce different approaches to estimation of misclassification parameters. Chapter 4 explores possible dependence between misclassification parameters in a binary dual test protocol in which we investigate the impact of model specification in marginally identifiable scenarios. Finally, in Chapter A we include the R and WinBUGS code for execution of the methods described in the chapters.

CHAPTER TWO

Bayesian Variable Selection for a Logistic Regression Model with a Misclassified Binary Covariate

2.1 Introduction

Variable selection is a specific type of model selection in which a parsimonious subset of covariates are identified as significant in the prediction of a response. Bayesian approaches to variable selection have been a rapidly advancing area of the statistics literature over the past 15 years. Though numerous approaches have been developed for variable selection, as of the time of publication, few have dealt with selecting a parsimonious set of covariates in the presence of misclassified data. We introduce a general approach to Bayesian variable selection for logistic regression models when a binary covariate is misclassified.

Model selection methods have a rich history in the classical statistical literature, with substantial contributions regarding how we relate information to statistical theory made by Kullback and Leibler (1951), Akaike (1973), and Mallows (1973). Most of the major gains in the Bayesian literature regarding model and variable selection have occurred since the early 1990s coinciding with the advent of computing power and subsequent widespread usage of Markov Chain Monte Carlo methods, including papers by Carlin and Chib (1995), George and McCulloch (1993), Kuo and Mallick (1998), Dellaportas et al. (2002), and Gelfand and Ghosh (1998), among others.

Variable selection simultaneously accomplishes several purposes in statistical inference. First, variable selection identifies the subset of covariates from a larger collection that significantly predicts changes or differences in the model outcome, thereby answering the investigational question of which covariates of a larger set satisfy some optimal parsimonious criteria. In epidemiological literature, model selection is frequently used to identify which covariates confound the relationship between some exposure and outcome (Greenland, 1989). Secondly, the resulting model accomplishes dimension reduction, selecting only the important covariates from a larger set for a more parsimonious model and thus minimizing the problems associated with having a large number of covariates relative to the sample size. Thirdly, most properly applied variable selection procedures should be able to eliminate potential multicollinearity in the model by identifying the covariate(s) most significantly related to the outcome and removing covariates that are linear combinations or near-linear combinations of the significant covariates.

Developing Bayesian approaches to variable selection poses a unique challenge. Typically there are numerous competing models, and typical decision-making tools such as Bayes factors are ill-equipped to handle the dimensionality of the problem. For a model with k covariates, there are 2^k possible competing models from which to choose. Moreover, as with any statistical inference, consideration must be made to prior knowledge of the problem, and great care must be exercised to produce prior probabilities that reflect what prior knowledge may exist without unwittingly inducing prior distributions contrary to the true nature of the parameters.

Misclassification of binary data in generalized linear models has been shown to produce biased estimates of the relationships between covariates and responses. In this chapter, we address the situation where a single binary covariate is potentially misclassified. For example, we may be forced to rely on a fallible diagnostic test to determine participants' disease status. The latent gold standard value is replaced by a fallible surrogate measure. Suppose a means to acquire a gold standard measurement exists but its use is prohibitively expensive or invasive for widespread use in the study. In such situations, researchers will often gather measurements based on the surrogate variable on all individuals and additionally obtain the gold standard measure for a subsample of the study. This *validation* subset is said to be internal, because it is gathered from within the actual study cohort. Studies that are *externally* validated use estimates of sensitivities and specificities of the surrogate measure from data sources other than the study of interest and are not addressed in the current paper.

In many situations it is reasonable to assume that the sensitivity and specificity of the surrogate variable may in some way depend on the covariates, leading to differential misclassification of the covariate value in question. For example, using body mass index to define obesity status has been shown to be unreliable among Asian populations and athletes (Wang et al., 1994). In order to allow for maximum flexibility and accuracy of the model, we propose a model where the sensitivity and specificity of the surrogate measure is modeled allowing for dependence on the remaining covariates.

The use of a validation subsample, also referred as double sampling, can be an efficient and accurate basis to correct for bias introduced by misclassification (Tenenbein, 1970). However, the selection of the validation subsample itself must be able to provide sufficient and representative information about the larger study sample. The subsample must then be acquired via a known, random mechanism such as a simple random sample or stratified simple random sample to ensure consistency with the rest of the sample and thus the population. Furthermore, the size of the validation subsample will necessarily be constrained by cost or other practical considerations, but the number of observations must be large enough to handle estimation of potentially numerous covariates.

In this paper, we attempt to merge the ideas of variable selection and correcting for misclassification in a single procedure. To date, there are few published methods that simultaneously select model covariates while adjusting for misclassification. Gerlach and Stamey (2007) and Powers et al. (2009) are among the few papers at this time, and these methods focus on misclassified outcomes rather than covariates. The proposed method offers a convenient way to identify the appropriate set of covariates across multiple models simultaneously. Furthermore, the method estimates three important relationships: the relationship between the covariates and desired response, the measurement of the true exposure in the presence of the remaining perfectly observed covariates, and the sensitivities and specificities of the surrogate variable. Additionally the sensitivities and specificities may also depend on the remaining covariates. In the methods section, we introduce the models and define the potential nature of the misclassification, followed by a discussion of the variable selection method we use. Next, we perform four simulations using the stated method on a set of randomly generated data. Then we apply the method to an existing data set to demonstrate its effectiveness in providing a parsimonious model while adjusting for misclassification. Finally, we discuss the results as well as considerations for future research.

2.2 Methods

2.2.1 Models

2.2.1.1 *Disease model.* We consider a typical logistic regression scenario from prospectively observed data. Suppose we wish to estimate the probability of observing a random outcome of interest for the ith individual, Y_i, such that $Y_i \sim Bernoulli(\pi_i)$. Furthermore, assume that for random outcome variable Y_i of our N independent study participants we observe the P-dimensional covariate vector $\mathbf{z}_i$ such that

$$logit(\pi_i) = logit(pr(Y_i = 1|\mathbf{z}_i)) = \beta_0 + \mathbf{z}_i'\boldsymbol{\beta}.$$

If all P covariate values as well as response Y are measured without misclassification, then

$$pr(Y_i = 1|\mathbf{z}_i) = \frac{exp(\beta_0 + \mathbf{z}_i'\boldsymbol{\beta})}{1 + exp(\beta_0 + \mathbf{z}_i'\boldsymbol{\beta})}$$

is consistent for π_i.

Let us now consider the case in which our response depends upon $\mathbf{z}$ and T. The model is

$$logit(pr(Y_i = 1|\mathbf{z}_i, T_i)) = \beta_0 + \mathbf{z}_i'\boldsymbol{\beta} + T_i\beta_{P+1},$$

but suppose T is a gold standard measure that is only available for a subset n of the larger study sample N. Instead, suppose we observe covariate X, a potentially misclassified binary exposure variable. For our model, we now have our response Y, which is related to perfectly observed covariates $\mathbf{z}$ and potentially misclassified covariate X. If we were to simply substitute X for T, our model would potentially be biased for π (Carroll et al., 2006).

Because both measures are binary, it is reasonable to assume that

$$T \sim Bernoulli(p)$$

and that

$$X \sim Bernoulli(q)$$

where

$$q = pr(X = 1|T, S, C) = T \times S + (1 - T) \times (1 - C),$$

and S and C are the sensitivity and specificity, respectively, of the surrogate exposure indicator X. In this situation, q is the conditional probability of observing $X = 1$ given the known true exposure status as well as the sensitivity and specificity. Because we are assuming our observational data are prospectively ascertained, rather than retrospectively as in a case-control study, it is reasonable to assign the simplifying assumption that the misclassification probabilities S and C are independent of the response but may depend on other covariates contained in $\mathbf{z}$. Violations of this assumption, when the rate of misclassification also depends on the value of the response such as in case-control studies, require additional levels of complexity, as described by Prescott and Garthwaite (2005).

We can subsequently model the outcome π by imputing the unobserved T from X through the validation subsample. This will require estimation of the measurement relationship between T and $\mathbf{z}$ as well as the dependence of the sensitivities and specificities on $\mathbf{z}$.

2.2.1.2 *Measurement model.* Typically in medical and epidemiological studies, the exposure of an individual depends on a number of demographic and environmental factors. This is unique to observational data due to the lack of researchers' control over assignment of the treatment or exposure of interest to the study participants (Cochran and Chambers, 1965). In our present case of a misclassified covariate with an available validation subsample, we are solely interested in the relationship between T and the remaining perfectly measured covariates, $\mathbf{z}$. As previously defined, $pr(T = 1) = p$, but now we are allowing more flexibility by defining the probability of exposure p as dependent on the other (perfectly observed) covariate values, $\mathbf{z}$. Thus, the probability of observing exposure T given the observed covariate vector $\mathbf{z}$ is

$$logit(pr(T_i = 1|\mathbf{z}_i)) = logit(p_i) = \lambda_0 + \mathbf{z}_i'\boldsymbol{\lambda}.$$

2.2.1.3 *Sensitivity and specificity models.* As previously mentioned, often it is a reasonable possibility that the sensitivity and specificities of diagnostic tests differ based on the observed covariates. For example, in studies where exposure to a disease is of interest, a test may have different detection rates between genders, or a disease state may be masked by comorbidities and confounding variables. In such a situation, it is important to adjust based on these covariates. Likewise, if exposure to an occupational hazard is of interest, often exposure may depend on the duration of employment, age, or the worker's physical location in relation to the hazard.

It is reasonable to assume then that the sensitivity of the test for the ith observation given covariate vector $\mathbf{z}$ can be modeled as

$$logit(S_i|\mathbf{z}_i) = logit(pr(X_i = 1|T_i = 1, \mathbf{z}_i)) = \gamma_0 + \mathbf{z}_i'\boldsymbol{\gamma} \tag{2.1}$$

and likewise the specificity as

$$logit(C_i|\mathbf{z}_i) = logit(pr(X_i = 0|T_i = 0, \mathbf{z}_i)) = \mu_0 + \mathbf{z}_i'\boldsymbol{\mu}. \tag{2.2}$$

Both the sensitivity and specificity are modeled as logistic responses and possibly depend on the P perfectly covariates contained in $\mathbf{z}$.

The posterior distribution for the model parameters without the variable selection component is shown below:

$$\begin{aligned} f(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\mu}|Y,X,\mathbf{z}_i,T_i) &\propto f(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\mu}) \\ &\times \prod_{i=1}^{N} pr(Y_i = 1|\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\mu},\mathbf{z}_i,X_i,T_i) \times \prod_{i=1}^{n} pr(T_i = 1|\boldsymbol{\lambda},\mathbf{z}_i) \\ &\times \prod_{i=1}^{n} pr(X_i = 1|\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{z}_i,T_i = 1) \times \prod_{i=1}^{n} pr(X_i = 0|\boldsymbol{\mu},\boldsymbol{\lambda},\mathbf{z}_i,T_i = 0) \end{aligned}$$

2.2.2 Bayesian Variable Selection

In the preceding models, we have identified four interrelated statistical models with potentially a large total number of covariates. It is unlikely that all the covariates that are important for estimating the sensitivity, S, are also important for estimating the probability of exposure, p, for example. The disease model as stated presently depends upon $P + 1$ covariates, and the measurement, sensitivity, and specificity models each depend upon P covariates. Under our considered four models, there are potentially $M = 4P + 1$ covariates that could be included for estimation of the outcome, measurement, and misclassification probabilities. To achieve parsimony, it is desirable to identify the covariates that significantly contribute to the model while excluding those that do not.

Using a similar argument as Dellaportas et al. (2002), suppose our response random variable Y is generated from model g, where g is one of G competing models under consideration. Then we define the set of parameters for model g by θ_g, where $\theta_g \in \Theta_g$ and Θ_g represents all possible values for the coefficients of model g. We recognize $f(y|g, \theta_g)$ defines the likelihood for the data under model g, and thus there are G possible discrete representations of $f(y|g, \theta_g)$. If we assign prior probability $f(g)$ to model g, then the posterior probability for model g can be expressed as

$$f(g|y) = \frac{f(g)f(y|g)}{\sum_{g \in G} f(g)f(y|g)}, g \in G.$$

We define the marginal likelihood for m as

$$f(y|g) = \int_{\theta_g \in \Theta_g} f(y|g, \theta_g) f(\theta_g|g) d\theta_g,$$

where $f(\theta_g|g)$ is the conditional prior distribution for the parameters θ_g.

Kuo and Mallick (1998) proposed placing independent binary indicator variables on covariates in general and generalized linear models. The vector of indicator variables that yields the highest posterior probability identifies the preferred subset of covariates based on the observed data. We extend this method to identify the significant subset of covariates across our four related logistic regression models to not only estimate the unbiased relationship between the response and the necessary perfectly observed covariates but also to estimate the important covariates governing the measurement, sensitivity, and specificity of the imperfectly measured covariate.

To this end, we define $\boldsymbol{\omega}' = (\omega_1, \omega_2, \ldots, \omega_M)$. This vector contains the M indicator variables for each of the parameters contained in the four models. We note that because each ω_j must take either 0 or 1 in value, there are 2^M possible manifestations of ω. Therefore, let us define $\boldsymbol{\omega}_k \in \Omega$, where Ω is the set of all possible 2^M vectors and $\boldsymbol{\omega}'_k = (\omega_{1k}, \omega_{2k}, \ldots, \omega_{Mk})$.

In previous sections we introduced the model parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}$, which specify the disease model, measurement model, sensitivity, and specificity, respectively. Let us define the M-dimensional vector $\boldsymbol{\theta}' = \begin{bmatrix} \boldsymbol{\beta} & \boldsymbol{\lambda} & \boldsymbol{\gamma} & \boldsymbol{\mu} \end{bmatrix}$ and thus

$$\vartheta' = (\boldsymbol{\omega}'\boldsymbol{\theta})' = (\beta_1\omega_1, \ldots, \beta_{P+1}\omega_{P+1}, \lambda_1\omega_{P+2}, \ldots,$$
$$\lambda_P\omega_{2P+1}, \gamma_1\omega_{2P+2}, \ldots, \gamma_P\omega_{3P+1}, \mu_1\omega_{3P+2}, \ldots, \mu_P\omega_M),$$

which contains all possible parameters, multiplied by the appropriate indicator variable from $\boldsymbol{\omega}$. We multiply each model parameter in θ by the corresponding element in $\boldsymbol{\omega}_k$ to obtain posterior estimates of elements of θ given $\boldsymbol{\omega}_k$.

The posterior distribution of the model adding the variable selection component appears as follows:

$$f(\boldsymbol{\omega}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\mu}) = f(\boldsymbol{\omega}|\boldsymbol{\theta}) \propto f(\boldsymbol{\omega}) \prod_{j=1}^{M} f(\boldsymbol{\theta}|\boldsymbol{\omega}).$$

We select

$$\max_{\boldsymbol{\omega}_k \in \Omega} pr(\boldsymbol{\omega}_k|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\mu})$$

as the most likely data-generating model.

The models from the preceding sections now appear as:

$$logit(pr(Y_i = 1|\boldsymbol{\beta}, \mathbf{z}_i, T_i, \boldsymbol{\omega})) = \beta_0 + \sum_{i=1}^{P} \beta_i Z_i \omega_i + \beta_{P+1} T_i \omega_{P+1},$$
$$logit(pr(T_i = 1|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{z}_i)) = \lambda_0 + \sum_{i=1}^{P} \lambda_i Z_i \omega_{\{P+1+i\}},$$
$$logit(pr(X_i = 1|\boldsymbol{\gamma}, \boldsymbol{\omega}, T_i = 1, \mathbf{z}_i)) = \gamma_0 + \sum_{i=1}^{P} \gamma_i Z_i \omega_{\{2P+1+i\}},$$
$$logit(pr(X_i = 0|\boldsymbol{\mu}, \boldsymbol{\omega}, T_i = 0, \mathbf{z}_i)) = \mu_0 + \sum_{i=1}^{P} \mu_i Z_i \omega_{\{3P+1+i\}}.$$

The biggest advantage of the proposed method is that it eliminates the need to compute multiple Bayes' factors or to invoke complicated multiple-dimensional

MCMC sampling schemes. In the present application the scheme allows for simultaneous estimation across the four models to obtain the most parsimonious set of covariates describing the overall relationship between the covariates, exposures, and disease, empirically estimated from the prior information and likelihood functions. Furthermore, the selection process is not only intuitively appealing, but it also satisfies a critical optimality property: the vector of covariates with the highest posterior probability minimizes the Bayes risk under zero-one loss (Kadane and Lazar, 2004).

2.2.2.1 *Prior elicitation and specification.* All necessary information for the likelihood of the model parameters, including the misclassification parameters, is contained in $\mathbf{z}$, X, Y, and the validation subsample data T. Thus, in the absence of expert opinion or prior information, we could use so-called "noninformative" priors of $N(\mathbf{0}, \sigma^2 \mathbf{I}_M)$ on $\boldsymbol{\theta}$ where $\mathbf{0}$ is an M-dimensional vector with each value equal to 0, $\mathbf{I}_M$ is the $M \times M$ identity matrix, and σ^2 is an arbitrarily large variance value that reflects one's prior knowledge about the variance of the parameter mean, such as 1000. Furthermore, independent diffuse normal priors, such as $N(0, 1000)$, could be placed on the model intercept terms.

Next, lacking prior evidence, we assumed that each ω_j is distributed as an independent $Bernoulli(\phi_j)$ random variable, and thus a reasonable prior value to reflect this uncertainty is $\phi_j = 0.5$. This induces the prior probability for the kth vector from Ω is $pr(\boldsymbol{\omega}_k) = 1/M$. Each covariate has equal probability of inclusion or exclusion, and the prior independence assumption of the covariates will almost certainly be lost in the posterior distributions.

2.2.2.2 *Computation.* We arrive at posterior density estimates of the parameters typically via Gibbs sampling. The described models can easily be implemented in a statistical software package capable of handling MCMC procedures such as Win-

BUGS. However, for those who desire more control over the specific sampling routine, the papers published by Kuo and Mallick (1998) and Dellaportas et al. (2002) discuss strategies for Gibbs sampling routines to arrive at posterior estimates, which can be implemented in computational packages such as R or MatLab. It should be noted, especially because these methods are developed with large data sets in mind they can be fairly time consuming to execute.

2.3 Simulation

To test the proposed method, we performed four simulations of varying sample sizes using pseudo-randomly generated values from known parametric distributions. We used the following combinations of total sample sizes (N) and validation sample sizes (n): 1) $N = 2000$, $n = 500$; 2) N=1000, n=500; 3) $N = 2000$, $n = 250$; and 4) $N = 1000$, $n = 250$. We defined the outcome variable Y to be related via the logit link to $p = 3$ covariates: Z_1, Z_2, and T, where T is a perfectly observed exposure that has been ascertained for n individuals. The surrogate value X is observed for all N participants. Using the n perfectly observed values, we can estimate the exposure using the estimated sensitivities and specificities for the remaining $N - n$ observations.

The sample sizes above were selected specifically to demonstrate the performance of the method under varying N and n combinations. The latter two examples, where $n = 250$, are especially informative about the ability of our method to select the "correct" parameter set (in the sense that the procedure identifies a model that includes the appropriate nonzero coefficients). Given our specified parameter values, the model in some circumstances will have too few observations in the validation sample for estimation of the sensitivities and specificities within certain subpopulations. These examples can be viewed as variable selection in the presence of inappropriately small validation sample sizes.

Initially we generated N observed values of $\mathbf{z} = (Z_1, Z_2)$ as independent Bernoulli random variables with probabilities 0.4 and 0.5, respectively. These represent the perfectly observed covariates. In the measurement model, we defined the true exposure, T, to depend on the covariates $\mathbf{z}_i$ via a logistic regression model with parameter values of $\lambda = (-0.2, 1.6, -2.0)$. Next, we define the sensitivity and specificity that describe the relationship between T and X according the the models above, using sensitivity parameter vector $\gamma = (2.5, 0, -1.8)$ and specificity parameter vector $\mu = (0.4, 1.6, 0)$. Thus, the sensitivity $pr(X = 1|T = 1)$ depends only on Z_2, and the specificity $pr(X = 0|T = 0)$ depends only on Z_1 in this hypothetical situation. Finally, we generate the response Y using the full unbiased logistic regression model with parameter values $\beta = (-1, 0, 1.1, -0.8)$.

The full code for use in WinBUGS v1.4.1 is included in the Appendix. We code four logistic regression models with all possible covariates (9), multiplying indicator variable ω_i, where $i = 1, \ldots, 9$, to the ith covariate across the models. The resulting model appears as

$$\begin{aligned}
logit(pr(Y|\boldsymbol{\beta}, \mathbf{z}, T)) &= \beta_0 + \beta_1 Z_1 \omega_1 + \beta_2 Z_2 \omega_2 + \beta_3 T \omega_3 \\
logit(pr(T = 1|\boldsymbol{\lambda}, \mathbf{z}, X)) &= \lambda_0 + \lambda_1 Z_1 \omega_4 + \lambda_2 Z_2 \omega_5 \\
logit(pr(X = 1|\boldsymbol{\gamma}, T = 1, \mathbf{z})) &= \gamma_0 + \gamma_1 Z_1 \omega_6 + \gamma_2 Z_2 \omega_7 \\
logit(pr(X = 0|\boldsymbol{\mu}, T = 0, \mathbf{z})) &= \mu_0 + \mu_1 Z_1 \omega_8 + \mu_2 Z_2 \omega_9
\end{aligned}$$

At this point it becomes apparent that, unlike other proposed variable selection procedures, it is only necessary to explicitly write the full model once in the code, and the algorithm then identifies the appropriate covariates to include, returning the posterior probability of each of the $2^9 = 512$ possible vectors of covariates.

We generate 150 data sets at each of the specified sample sizes and obtained the results that follow. Note, for the sake of space, we only included information that indicates which model was selected with the highest posterior probability as well as the rank of the correct (data generating) model. Each model was enumerated from 1 to 512, and the data generating model is identified as model 223.

All data were generated by R version 2.5.1. Using the R2WinBUGS package (Gelman et al., 2004), the data passed into WinBUGS version 1.4.1 where all posterior distributions were estimated via MCMC sampling. For each generated dataset, we executed 3 chains with different initial conditions. We discarded a 5000 iteration burn-in to allow the chains to converge to the target distribution and then we kept a 15,000 iteration sample of the posterior distribution. We observed no major issues with convergence or autocorrelation in the resulting chains. The decision to execute 150 simulations per sample size was mainly related to the time required to execute a single iteration of the model, which required up to 45 minutes depending on the sample size and processor. Furthermore, the sequential nature of WinBUGS limits our ability to fully utilize computing advances such as multiple-core processors.

2.3.1 Simulation Results for N=2000, n=500

The first simulation has a sufficiently large sample size to be able to properly identify the subset of model covariates of the highest posterior probability. In Figure 2.1 we observe that the algorithm identifies the appropriate set of covariates with observed probability $140/150 = 0.93$. The next highest frequency model, number 219, contains all nonzero coefficients except β_3. Its value ($\beta_3 = -0.8$) has the smallest absolute distance from zero of the elements of the β-vector, and thus this would be the one least likely to be detected. The algorithm omits β_3 with observed relative frequency $6/150 = 0.04$. Finally, for the remaining 4 generated datasets, the model selection procedure identifies model 159, which contains all appropriate nonzero covariates except for $\gamma_2 = -1.8$. This covariate aids in estimating the sensitivity ,and all information that contributes to the posterior mean of λ, γ, or μ comes from the validation subsample.

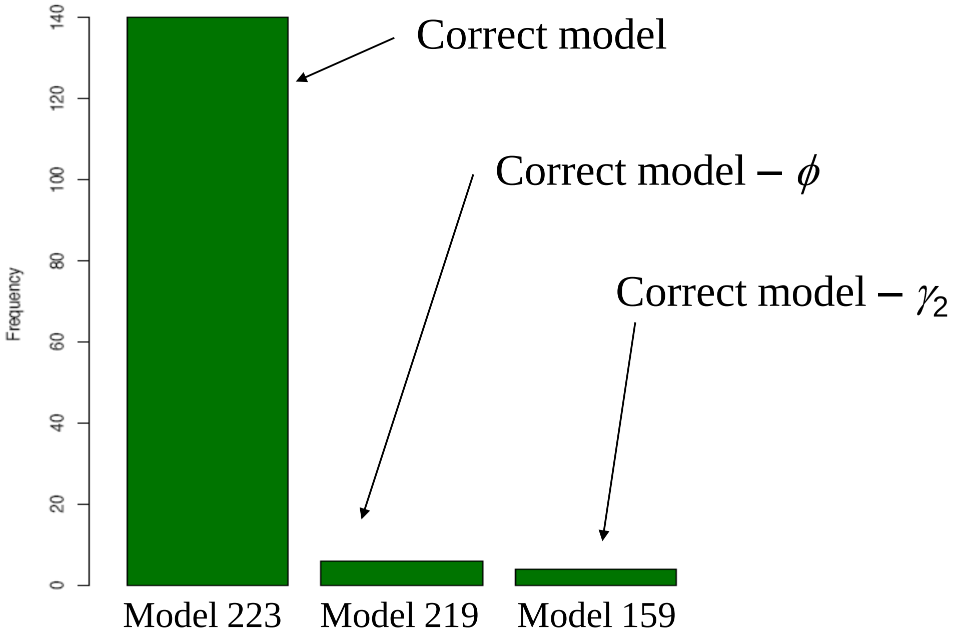

Figure 2.1: Frequencies of posterior highest posterior probability models for 150 simulated data sets; N=2000, n=500.

2.3.2 Simulation Results for N=1000, n=500

The objective of the second simulation is to assess the impact of reducing the overall sample size, N, while maintaining the same validation subsample size, n. In Figure 2.2, we observe that the algorithm appropriately identifies model 223 as having the highest posterior probability 115 times of the 150 simulated data sets for an estimated probability of 0.77. We note that among the competing models, model 219 (described in the previous paragraph) is the second highest frequency model, selected 29 times. Model 159 (described previously) and model 155 (same as 223 except omits β_3 and γ_2) were each chosen 3 times.

The finding of note is that when we decrease the overall sample size N, we impair the performance of the model selection procedure. However, the results are still overwhelmingly favorable under the given parameter values. The overall reduction in the sample size still selects the correct model with the highest frequency. The biggest difference occurs in the method's ability to correctly include β_3, where $32/150 = 0.21$ of the selected models omit the covariate. However, we maintained the validation subsample size of $n = 500$, and thus the observed probability of selecting a model that incorrectly omits γ_2 remained $6/150 = 0.04$, as in the first simulation. This is noteworthy because the estimation of γ_2 can only be done from the validation data. Thus, we see the reduction in the overall sample size changed our ability to identify a covariate from the disease model, but the models that depend on the validation data set (i.e., any model that requires exact knowledge of T) remain unaltered by this reduction.

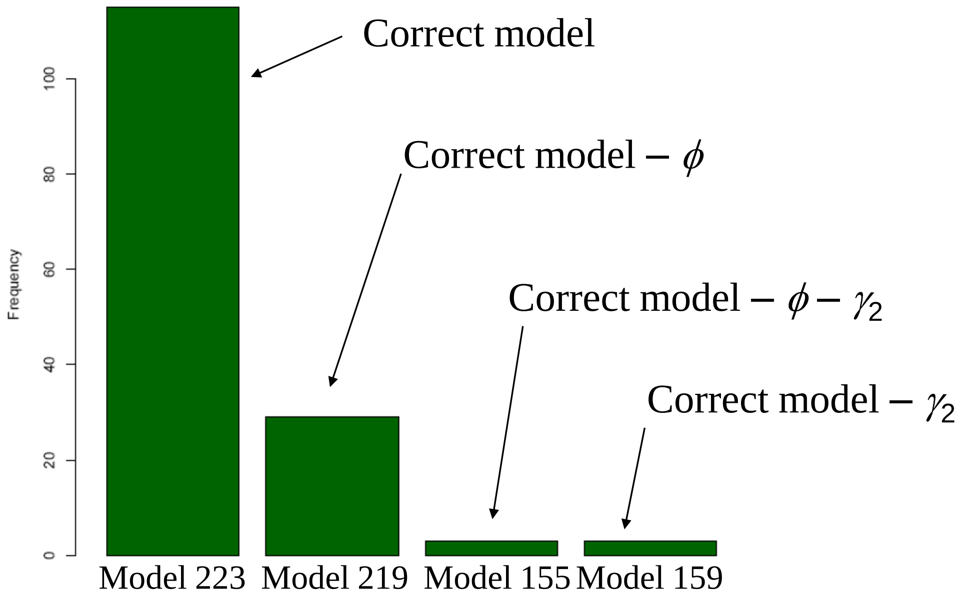

Figure 2.2: Frequencies of posterior highest posterior probability models for 150 simulated data sets; N=1000, n=500.

2.3.3 Simulation Results for N=2000, n=250

For the next two simulations, we reduce the validation subset size by 50% and evaluate the model selection procedure's performance. In Figure 2.3, we see that the correct model, 223, is still identified as the highest posterior model a majority of the time and that it far outperforms any competing model. However, the excluded models again provide information on what portions of the model are impaired by the reduction of the subsample size; the second most frequently selected highest posterior probability model is that of 159, the model that contains all appropriate covariates except that it excludes γ_2. This is noteworthy because the reduction in the validation sample size limits our ability to estimate and thus include covariates that help measure the relationship between the misclassified X and gold-standard T.

Less frequently, the procedure selects the model that omits the β_3 covariate, and in an even smaller number of cases omits both β_3 and γ_2. We do note that the frequency with which the β_3 covariate is excluded is higher in this case than in the first simulation where $N = 2000$ and $n = 500$. We observe that simply having a large number of observations doesn't guarantee proper estimation and performance of the model selection procedure; having sufficient gold-standard validation information strongly aids in our ability to estimate the disease model covariates with certainty.

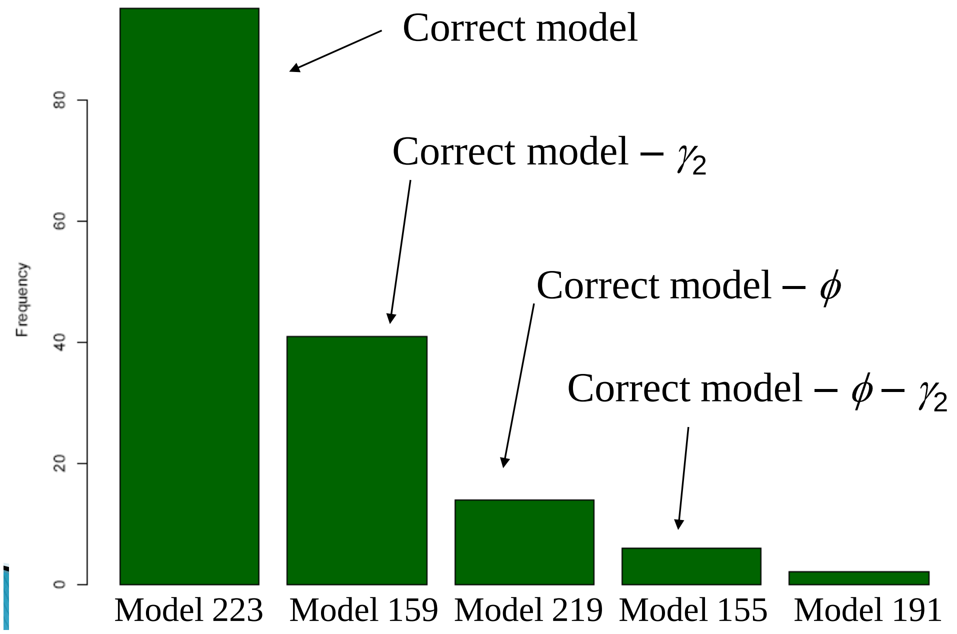

Figure 2.3: Frequencies of posterior highest posterior probability models for 150 simulated data sets; N=2000, n=250.

2.3.4 Simulation Results for N=1000, n=250

The fourth and final simulation evaluates the variable selection procedure's performance under minimal overall sample size and validation subsample size. At this point, we have observed that the variables that are most severely affected by such reductions appear to be β_3 and γ_2. As we see in Figure 2.4, this pattern continues. For the first time, we have identified a scenario in which the true data generating model is correctly identified less than half the time, but the correct model is still assigned with the highest posterior density more frequently than any other model. Under these conditions, the models 219 and 159 appear to be selected with similar frequency, indicating that when model 223 is not selected, the procedure typically fails to identify only one covariate. Slightly less than 10% of the time, the procedure identifies a model that excludes both covariates.

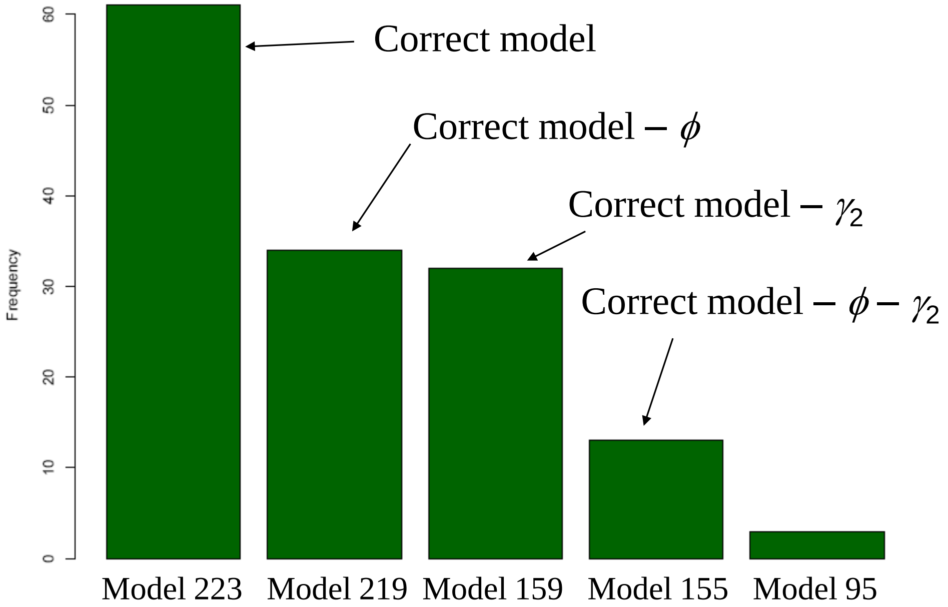

Figure 2.4: Frequencies of posterior highest posterior probability models for 150 simulated data sets; N=1000, n=250.

2.4 Example

To demonstrate the method on previously published data, we use data from highway safety research published by Hochberg (1977). In this study, we wish to determine whether the probability of an automobile accident producing an injury depends on the binary covariates driver gender, car damage (low or high), and seat belt use. However, the information on the police reports regarding belt usage is subject to systematic misclassification errors that may potentially depend on the other covariates gender and car damage. We assume all other data contained on the police report are considered accurate. The majority of the data observations $(80,084)$ have data recorded solely from police accident reports, while a subsample $(n = 1,796)$ of drivers had the police accident report along with a more intensive follow-up interview that contains what we will consider gold standard information regarding seat belt usage. The total sample size is $N = 81,880$.

Initially, we analyze the data ignoring misclassification. We define the response Y_i as the binary response of injury status (0 = uninjured, 1 = injured) for the ith individual $(i = 1, \ldots, N)$. Furthermore, the fallible measure for belt usage is denoted by X_i, where $X_i = 1$ indicates that belt usage was reported by the police and $X_i = 0$ indicates no reported belt usage. The indicators $Z_{1,i}$ and $Z_{2,i}$ represent gender (1=male) and car damage (1=high), respectively. The Bernoulli response Y_i occurs with probability π_{iN}, where

$$logit(\pi_{iN}) = \beta_{0N} + \beta_{1N} X_i + \beta_{2N} Z_{1,i} + \beta_{3N} Z_{2,i}.$$

In this formulation, $\boldsymbol{\beta}_N$ is the parameter vector for the naïve model that ignores misclassification. We subsequently multiply each element of $\boldsymbol{\beta}_N$ by a Bernoulli indicator variable so that our model becomes

$$logit(\pi_{iN}) = \beta_{0N} + \omega_1 \beta_{1N} X_i + \omega_2 \beta_{2N} Z_{1,i} + \omega_3 \beta_{3N} Z_{2,i}. \tag{2.3}$$

Because there are three parameters in the model to consider, the procedure must choose between $2^3 = 8$ possible models. We place diffuse normal prior distributions on the model parameters such that $\boldsymbol{\beta}_N \sim N(0, \sigma^2 \mathbf{I}_4)$, where $\sigma^2 = 1,000$ and $\mathbf{I}_j$ is the $j \times j$ identity matrix. Furthermore, we place independent $Bernoulli(0.5)$ prior distributions on each ω_k where $k = 1, 2, 3$.

To account for the misclassified covariate X, we define the true belt usage indicator variable T_i for the ith individual where $T_i = 1$ indicates the study participant was wearing a belt and $T_i = 0$ if not. The variable T is observed for only the subsample ($n = 1,796$) and is distributed Bernoulli with probability p. We now define the Bernoulli response Y_i as occurring with probability π_i, where

$$logit(\pi_i) = \beta_0 + \beta_1 T_i + \beta_2 Z_{1,i} + \beta_3 Z_{2,i}.$$

Furthermore, we allow the probability of T ($pr(T = 1) = p$) as well as the sensitivity (S) and specificity (C) of X to depend on Z_1 and Z_2 as follows:

$$\begin{aligned} logit(p_i) &= \lambda_0 + \lambda_1 Z_{1,i} + \lambda_2 Z_{2,i}, \\ logit(S_i) &= \gamma_0 + \gamma_1 Z_{1,i} + \gamma_2 Z_{2,i}, \\ logit(C_i) &= \mu_0 + \mu_1 Z_{1,i} + \mu_2 Z_{2,i}. \end{aligned}$$

To develop a more parsimonious model, we multiply binary indicator parameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_9)$ to each covariate of the set of models so that our system of equations becomes

$$logit(\pi_i) = \beta_0 + \omega_1\beta_1 T_i + \omega_2\beta_2 Z_{1,i} + \omega_3\beta_3 Z_{2,i}, \tag{2.4}$$

$$logit(p_i) = \lambda_0 + \omega_4\lambda_1 Z_{1,i} + \omega_5\lambda_2 Z_{2,i}, \tag{2.5}$$

$$logit(S_i) = \gamma_0 + \omega_6\gamma_1 Z_{1,i} + \omega_7\gamma_2 Z_{2,i}, \tag{2.6}$$

$$logit(C_i) = \mu_0 + \omega_8\mu_1 Z_{1,i} + \omega_9\mu_2 Z_{2,i}. \tag{2.7}$$

Assuming no prior information we place diffuse normal prior distributions for each of the model parameters in the manner of

$$\begin{aligned} \boldsymbol{\beta} &\sim N(0, \sigma^2\mathbf{I}_4), \\ \boldsymbol{\lambda} &\sim N(0, \sigma^2\mathbf{I}_3), \\ \boldsymbol{\gamma} &\sim N(0, \sigma^2\mathbf{I}_3), \\ \boldsymbol{\mu} &\sim N(0, \sigma^2\mathbf{I}_3). \end{aligned}$$

Table 2.1. Top Five Posterior Model Probabilities for Naive Model

Covariate Vector	Posterior probability
$(\beta_{1N}, \beta_{2N}, \beta_{3N})$	0.996
$(\beta_{1N}, 0, \beta_{3N})$	0.004
All others	0

Next, for each ω_k $(k = 1, \ldots, 9)$, we assign prior distribution

$$\omega_k \sim Bernoulli(0.5).$$

In this situation we have $2^9 = 512$ possible models with an equal induced prior probability for each model. We assign initial values to each of three chains, execute a $5,000$ iteration burn-in, and then keep $15,000$ samples from the posterior distribution for each chain. The five highest probability posterior distributions are listed in Table 2.3.

We first observe in Table 2.4 the posterior model probabilities for the variable selection procedure from model 2.3 in which we naïvely assume X is observed perfectly. It should come as no surprise that, of the eight possible models, overwhelmingly the highest posterior model includes all three binary covariates. Given the extremely large sample size $N = 81,880$, all parameter standard deviation estimates are small enough that the posterior mean supports $\omega_j = 1$ for $j = 1, 2, 3$.

In Table 2.4 we observe the parameter estimates for the naïve model. These shall be used as a basis for comparison with the posterior values from Table 2.4. For now, we simply note the model parameters are nonzero with small standard deviations, which is consistent with the general notion of a variable being useful for posterior prediction.

We observe in Table 2.3 the models identified by their posterior probability when we attempt to correct for misclassification. The highest posterior probability

Table 2.2. Highest posterior probability model parameter estimates

Parameter	Mean	Standard Deviation	95% Credible Interval
β_{0N}	-2.13	0.02	$(-2.18, -2.10)$
β_{1N}	1.54	0.02	$(1.50, 1.59)$
β_{2N}	-0.38	0.02	$(-0.42, -0.35)$
β_{3N}	-0.28	0.033	$(-0.35, -0.22)$

Table 2.3. Top 5 Posterior Model Probabilities for Misclassification Model

Covariate Vector	Posterior probability
$(\beta_1, \beta_2, \beta_3, 0, 0, 0, \gamma_2, \mu_1, 0)$	0.54
$(\beta_1, \beta_2, \beta_3, 0, \lambda_2, 0, 0, \mu_1, 0)$	0.24
$(\beta_1, \beta_2, \beta_3, \lambda_1, \lambda_2, 0, 0, 0, 0)$	0.19
$(\beta_1, \beta_2, \beta_3, \lambda_1, \lambda_2, 0, 0, \mu_1, 0)$	0.015
$(\beta_1, \beta_2, \beta_3, 0, \lambda_2, 0, \gamma_2, \mu_1, 0)$	0.0059

model has no covariates from the $\boldsymbol{\lambda}$ model, indicating that neither gender nor crash severity likely contribute significantly to the estimation of p.

Given this information, we can now construct the posterior distribution with the least posterior predictive variability. We fit the following model

$$\begin{aligned}
&logit(\pi_i) = \beta_0 + \beta_1 T_i + \beta_2 Z_{1,i} + \beta_3 Z_{2,i},\\
&logit(p_i) = \lambda_0,\\
&logit(S_i) = \gamma_0 + \gamma_2 Z_{2,i},\\
&logit(C_i) = \mu_0 + \mu_1 Z_{1,i}.
\end{aligned}$$

which yields the posterior estimates listed on Table 2.4.

We note that the posterior mean estimates for $\boldsymbol{\beta}$ in Table 2.4 have similar means to the posterior estimates for $\boldsymbol{\beta}_N$ in Table 2.4 with the notable exception for β_1 for T compared to β_1 for X. It appears that the misclassification resulted in an attenuation of the effect of seat belt usage on the probability of injury. When the misclassification has been accounted, we observe an even lower probability of injury for belt users than was observed by the misclassified parameter. Furthermore, the corrected model has a slightly larger parameter standard deviation, which reflects our added uncertainty due to the misclassified covariate.

2.5 Discussion

It should be noted from the preceding simulation results that this method of variable selection consistently identifies the “correct” model out of a large number of equally likely (based on assumed *a priori* probabilities) competing models a very high proportion of the time, even in the cases where the validation sample size is small relative to the overall study sample. The priors used for variable selection in this article extend the method proposed by Kuo and Mallick (1998) to misclassified data, and more generally, to simultaneous approximations of posterior distributions of multiple interrelated models.

Table 2.4. Highest posterior probability model parameter estimates

Parameter	Mean	Standard Deviation	95% Credible Interval
β_0	-2.11	0.02	$(-2.15, -2.06)$
β_1	1.54	0.02	$(1.51, 1.59)$
β_2	-0.39	0.02	$(-0.43, -0.35)$
β_3	-0.38	0.05	$(-0.47, -0.29)$
λ_0	-1.60	0.05	$(-1.70, -1.50)$
γ_0	-0.18	0.09	$(-0.34, -0.00)$
γ_2	0.37	0.05	$(0.26, 0.48)$
μ_0	3.58	0.14	$(3.33, 3.86)$
μ_1	0.87	0.20	$(0.53, 1.30)$

As mentioned previously, upon selecting the parameter vector of the highest posterior probability, we would now consider this model as being most likely the "data generating" vector, that is, the vector including the covariates that most likely generated the observed sample. In situations where the selected vector defies logic or does not clearly distinguish itself as the superior model, logic demands that we consider the next closest competitor or competitors as potential data generating models. While the selected model is optimal under the Bayes decision rule using zero-one loss, in practice closely competing models may force additional consideration for the researcher or in some cases selection of multiple *candidate* models. Although this situation is undesirable from a decision-making framework, often the observed data does not implicate a single vector of covariates as overwhelmingly being clearly the most probable, and thus the procedure should not be viewed with the same simplicity that is commonly reserved for misinterpretations of p-values. Besides, a variable selection procedure is more appropriately viewed as an investigational tool, and paring the possible models, for example, from over 500 to 3 or fewer can still be considered successful in the sense of how many possible models were removed from consideration.

The simulations in this paper were designed mainly to correspond with the example cited. Thus, this paper uses binary data as both covariates and responses. Future work could include model selection for models with continuous covariates. Kuo and Mallick (1998) successfully introduced their method using continuous covariates, so it is unlikely that any major differences will result. The appeal of the proposed method is its ease in implementing over a complicated model. The decision to select only covariates and not intercept terms of the sensitivity, specificity, and measurement models was based on the knowledge that the binary variable was misclassified. The process of selecting the covariates can be viewed as letting the data determine whether the misclassification was *differential*, i.e. the sensitivities, specificities, or measurement differed with respect to the other covariates. A model that excludes all covariates for the sensitivities and specificities would indicate that the misclassification is *nondifferential*.

CHAPTER THREE

Bayesian Approaches to Pass-Fail Testing for Biostatistical Applications

3.1 Introduction

Repeated binary testing (i.e. pass-fail inspections, go/no-go testing) is regularly utilized in quality control studies as a means to classify items when quantitative methods are prohibitive or nonexistent (Boyles, 2001). Biostatisticians have adapted several of these developments for application in biological studies, modifying such methods to be able to account for the complex variability present in biological systems (Fujisawa and Izumi, 2000). Fallible classification systems introduce bias into the measurement process, and reliable estimation of misclassification probabilities helps yield individual classification and population prevalence estimates with greater precision and consistency.

Few studies have utilized the strength of the Bayesian approach in the development of methods to estimate parameters in binary classification systems. Bayesian estimation through Markov Chain Monte Carlo simultaneously estimates classification probabilities at the individual and population level, and the possibility of including prior information remains a strength of the Bayesian approach. Hierarchical statistical modeling is another option made readily available through Bayesian estimation that is especially applicable when the data arise from biostatistical data.

The intent of this work is to develop a Bayesian framework to develop a general approach to parameter estimation for pass-fail models, especially within the context of biostatistical approaches for detecting the presence of biomarkers. This includes the development of MCMC approaches to sampling from the posterior distributions of the misclassification probabilities as well as the construction of hierarchical approaches to account for potentially unmeasured interindividual variability. In quality control settings, it is often reasonable to assume that individual items are homogeneous with regard to the probability of having “passable” qualities, and furthermore that the probability of misclassification by the inspection system is likewise homogeneous. In biostatistical studies, flexible methods must be readily available to account for heterogeneity among individuals that affect the ability to detect a disease state or exposure captured via a discretized biomarker.

In the methods section we propose various models for estimating assumed stationary misclassification probabilities. Furthermore, we propose a Gibbs sampler that can produce the desired posterior density estimates for the model parameters. In the following section we propose models for hierarchical estimation of misclassification probabilities where we assume the probability of misclassification for each individual arises hierarchically from a common distribution. We present results for a simulation to determine the necessary sample sizes to be able to produce reliable estimates of both prevalence and misclassification probabilities, and we follow this with examples using previously published data sets. Finally, we conclude by discussing future directions for subsequent research.

3.2 Modeling

For a sample of N randomly selected observations, we denote the true status of the observation with random variable T, where $T \sim Bernoulli(\tau)$. Thus, each observation contains the condition of interest with probability $\tau = Pr(T = 1)$, which will also be referred to as the prevalence of T. However, we assume that the measurement of the status uses a fallible classifier, and consequently the random variable T is latent.

Instead, we observe a repeated number of independent observations from the ith individual to assess the presence of the status, T_i, assumed to be stationary with respect to the observational interval. Let binary $Y_{i,j}$ denote the result of the jth time ($j = 1, 2, \ldots, n_i$) that individual i ($i = 1, 2, \ldots, N$) is tested. For each $Y_{i,j}$ we define the conditional probabilities $\theta_- = Pr(Y_{i,j} = 0|T_i = 1)$ (false negative rate) and $\theta_+ = Pr(Y_{i,j} = 1|T_i = 0)$ (false positive rate) with respect to latent T_i. The conditional densities of $Y_{i,j}$ are

$$Y_{i,j}|T_i = 0 \sim Bernoulli(1 - \theta_-),$$
$$Y_{i,j}|T_i = 1 \sim Bernoulli(\theta_+).$$

Note, an alternate definition of the misclassification probabilities could involve using the sensitivity $(1 - \theta_-)$ and specificity $(1 - \theta_+)$ of the measurement system, although the two approaches yield the same inference of the posterior distribution.

Finally, we define $X_i = \sum_{j=1}^{n_i} Y_{i,j}$. Because each of the n_i repeated tests are independent, the conditional distribution $X_i|T_i = 1$ is $binomial(n_i, \theta_+)$, and likewise $X_i|T_i = 0 \sim binomial(n_i, 1 - \theta_+)$. However, because T is latent, inference about the parameters depends on the joint distribution of X and T. A useful identity is

$$f(x_i, t_i|\theta_-, \theta_+, \tau) = f(x_i|t_i, \theta_-, \theta_+)f(t_i|\tau).$$

Hence the joint distribution of the observed data vector $\mathbf{x}$ and the latent vector $\mathbf{t}$ is

$$\begin{aligned} f(\mathbf{x}, \mathbf{t}|\theta_-, \theta_+, \tau) &= \prod_{i=1}^{N} f(x_i|t_i, \theta_-, \theta_+)f(t_i|\tau) \\ &\propto \prod_{i=1}^{N} \left(\tau(1 - \theta_-)^{x_i}(\theta_-)^{n_i - x_i}\right)^{t_i} \\ &\times \left((1 - \tau)(\theta_+)^{x_i}(1 - \theta_+)^{n_i - x_i}\right)^{1 - t_i}. \end{aligned}$$

3.3 Conjugate Beta Priors

Because θ_-, θ_+, and τ are probabilities and thus can be nonzero only within the unit interval, independent beta prior distributions are reasonable in the absence of other evidence (Joseph et al., 1995). Note that this parameterization could allow for a nonidentifiable model without the restriction $\theta_- + \theta_+ \geq 1$ (Rogan and Gladen, 1978). Let us proceed assuming independent prior distributions such that

$$\begin{aligned}
\theta_- &\sim beta(\alpha_-, \beta_-),\\
\theta_+ &\sim beta(\alpha_+, \beta_+),\\
\tau &\sim beta(\alpha_0, \beta_0).
\end{aligned}$$

Thus, the joint posterior distribution of the parameters is

$$\begin{aligned}
f(\theta_-,&\theta_+, \tau|\mathbf{x}, \mathbf{t}) \propto p(\theta_-, \theta_+, \tau) f(\mathbf{x}, \mathbf{t}|\theta_-, \theta_+, \tau)\\
&\propto \tau^{\sum_{i=1}^{N} t_i + \alpha_0 - 1}(1-\tau)^{N - \sum_{i=1}^{N} t_i + \beta_0 - 1}\\
&\times (\theta_-)^{\sum_{i=1}^{N} t_i(n_i - x_i) + \alpha_- - 1}(1-\theta_-)^{\sum_{i=1}^{N} t_i x_i + \beta_- - 1}\\
&\times (\theta_+)^{\sum_{i=1}^{N}(1-t_i)x_i + \alpha_+ - 1}(1-\theta_+)^{\sum_{i=1}^{N}(1-t_i)(n_i - x_i) + \beta_+ - 1}.
\end{aligned} \tag{3.1}$$

In the absence of any prior information, a common noninformative prior for each of θ_-, θ_+, and τ is the beta(1,1) distribution, which assumes equal probability for all possible values across the domain of each of the parameters. We observe

$$\begin{aligned}
f(\theta_-,&\theta_+, \tau|\mathbf{x}, \mathbf{t}) \propto p(\theta_-, \theta_+, \tau) f(\mathbf{x}, \mathbf{t}|\theta_-, \theta_+, \tau)\\
&\propto \tau^{\sum_{i=1}^{N} t_i}(1-\tau)^{N - \sum_{i=1}^{N} t_i}\\
&\times (\theta_-)^{\sum_{i=1}^{N} t_i(n_i - x_i)}(1-\theta_-)^{\sum_{i=1}^{N} t_i x_i}\\
&\times (\theta_+)^{\sum_{i=1}^{N}(1-t_i)x_i}(1-\theta_+)^{\sum_{i=1}^{N}(1-t_i)(n_i - x_i)}.
\end{aligned} \tag{3.2}$$

The posterior distribution shown in 3.2 takes the form of the frequentist likelihood function previously published (van Wieringen and van den Heuvel, 2005). However, we proceed under the general formulation of the posterior distribution shown in 3.1.

3.3.1 MCMC Sampling from the Posterior

Prior frequentist approaches to this problem have been proposed. Espeland and Hui (1987) discusses the use of latent class analysis for estimating parameters in the presence of binary misclassification. The approach of Reese et al. (2008) defines an E-M algorithm that produces point estimates somewhat comparable to the Gibbs Sampling approach described below. However, our approach is fully Bayesian and thus allows for the incorporation of prior information.

It is straightforward to sample from the posterior distribution using the general beta priors from (3.1).

i Begin with $n_1, n_2, \ldots, n_N$, our data vector $\mathbf{x}$. Set initial values $\tau^{(0)}$,$\theta_-^{(0)}$, and $\theta_+^{(0)}$.

ii Randomly generate data vector $\mathbf{t}^{(0)}$ of length N where $t_i^{(0)}$ is distributed $Bernoulli(\tau^{(0)})$.

iii For $k = 1, 2, \ldots$

(a) sample $\theta_-^{(k)}$ from $beta\left(\sum_{i=1}^{N} t_i^{(k)}(n_i - x_i) + \alpha_-, \sum_{i=1}^{N} t_i^{(k)} x_i + \beta_-\right)$.

(b) sample $\theta_+^{(k)}$ from $beta\left(\alpha_+ + \sum_{i=1}^{N}(1 - t_i^{(k)})x_i, \beta_+ + \sum_{i=1}^{N}\left(1 - t_i^{(k)}\right)(n_i - x_i)\right)$.

(c) sample $\tau^{(k)}$ from $beta\left(\sum_{i=1}^{N} t_i^{(k)} + \alpha_0, N - \sum_{i=1}^{N} t_i^{(k)} + \beta_0\right)$.

(d) generate $\mathbf{t}^{(k)}$ where $t_i^{(k)}$ is distributed $Bernoulli\left(\xi_i^{(k)}\right)$ and

$$\xi_i^{(k)} = \frac{\tau^{(k)}\left(\theta_-^{(k)}\right)^{n_i - x_i}\left(1 - \theta_-^{(k)}\right)^{x_i}}{\tau^{(k)}\left(\theta_-^{(k)}\right)^{n_i - x_i}\left(1 - \theta_-^{(k)}\right)^{x_i} + (1 - \tau^{(k)})\left(\theta_+^{(k)}\right)^{n_i - x_i}\left(1 - \theta_+^{(k)}\right)^{x_i}}.$$

Note:

$$\begin{aligned}
\xi_i^{(k)} &= Pr\left(T_i = 1 | X_i = x_i, \tau^{(k)}, \theta_-^{(k)}, \theta_+^{(k)}\right) \\
&= \frac{Pr\left(T_i = 1, X_i = x_i | \tau^{(k)}, \theta_-^{(k)}, \theta_+^{(k)}\right)}{Pr\left(X_i = x_i | \tau^{(k)}, \theta_-^{(k)}, \theta_+^{(k)}\right)} \\
&= \frac{Pr\left(T_i = 1, X_i = x_i | \tau^{(k)}, \theta_-^{(k)}, \theta_+^{(k)}\right)}{\sum_{t_i=0}^{1} Pr\left(T_i = t_i, X_i = x_i | \tau^{(k)}, \theta_-^{(k)}, \theta_+^{(k)}\right)}
\end{aligned}$$

(e) Store the parameter values of interest, increment k by 1, return to step (a), and repeat.

iv Burn in at least several hundred iterations of the Gibbs sampler until convergence to the posterior has been achieved, discard these initial values, and then keep several thousand iterations of each parameter to represent the full posterior distribution.

3.4 Alternate Parameterizations for Misclassification Probabilities

The preceding sections define the binary test model taking advantage of the relationships between the beta and binomial distributions. We now consider alternate means to define the model and estimate the parameters under a more diverse set of assumptions.

3.4.1 The Dirichlet Misclassification Model

The method of estimation of prevalence and misclassification probabilities previously described can be criticized as misstating the true domain of θ_- and θ_+. If θ_- and θ_+ are small, modifying the domain is unnecessary in practice. However, intuitively, it may be desirable to place a Dirichlet prior distribution on the misclassification parameters so that their shared domain is within the unit interval. We define

$$p(\theta_-, \theta_+) = \frac{1}{B(\alpha_1, \alpha_2, \alpha_3)} \theta_-^{\alpha_1 - 1} \theta_+^{\alpha_2 - 1} (1 - \theta_- - \theta_+)^{\alpha_3 - 1} \tag{3.3}$$

where $B(\alpha_1, \alpha_2, \alpha_3)$ is the beta function for $\alpha_1, \alpha_2, \alpha_3 > 0$, and $0 \leq \theta_- \leq 1$, $0 \leq \theta_+ \leq 1$, and $0 \leq \theta_- + \theta_+ \leq 1$. The added intuitive benefit of the Dirichlet distribution is that the false positive and false negative rates can be perceived as arising from the same random processes whose domains are dependent. Furthermore, a single Dirichlet prior distribution depends upon three hyperparameters rather than two independent beta distributions which rely on four hyperparameters, resulting in a moderate step toward increased parsimony.

We construct a joint prior distribution with misclassification parameters independent of τ, and we obtain the posterior distribution

$$\begin{aligned}f(\theta_-,&\theta_+,\tau|\mathbf{x},\mathbf{t}) \propto p(\theta_-,\theta_+,\tau)f(\mathbf{x},\mathbf{t}|\theta_-,\theta_+,\tau)\\ &\propto \tau^{\sum_{i=1}^N t_i+\alpha_0-1}(\theta_-)^{\sum_{i=1}^N t_i(n_i-x_i)+\alpha_1-1}(1-\theta_-)^{\sum_{i=1}^N t_i x_i}\\ &(1-\tau)^{\sum_{i=1}^N(1-t_i)+\beta_0-1}\times(\theta_+)^{\sum_{i=1}^N(1-t_i)x_i+\alpha_2-1}(1-\theta_+)^{\sum_{i=1}^N(1-t_i)(n_i-x_i)}\\ &(1-\theta_+-\theta_-)^{\alpha_3-1}.\end{aligned}$$

Sampling from the posterior distribution is similar to sampling from (3.1) with the exception that the first two steps are modified such that we jointly sample $(\theta_-^{(k)},\theta_+^{(k)})$ from the density

$$\begin{aligned}f(\theta_-^{(k)},&\theta_+^{(k)}|\mathbf{x},\mathbf{t}^{(k)}) \propto \left(\theta_-^{(k)}\right)^{\sum_{i=1}^N t_i^{(k)}(n_i-x_i)+\alpha_1-1}\left(1-\theta_-^{(k)}\right)^{\sum_{i=1}^N t_i^{(k)}x_i}\\ &\times\left(\theta_+^{(k)}\right)^{\sum_{i=1}^N(1-t_i^{(k)})x_i+\alpha_2-1}\left(1-\theta_+^{(k)}\right)^{\sum_{i=1}^N(1-t_i^{(k)})(n_i-x_i)}\\ &\times\left(1-\theta_+^{(k)}-\theta_-^{(k)}\right)^{\alpha_3-1}.\end{aligned}$$

These parameters can be sampled jointly from the unnamed distribution above, or each may be sampled separately from the conditional densities

$$\begin{aligned}f(\theta_-^{(k)}|\theta_+^{(k)},\mathbf{x},\mathbf{t}^{(k)}) &\propto \left(\theta_-^{(k)}\right)^{\sum_{i=1}^N t_i^{(k)}(n_i-x_i)+\alpha_1-1}\left(1-\theta_-^{(k)}\right)^{\sum_{i=1}^N t_i^{(k)}x_i}\\ &\times(1-\theta_+^{(k)}-\theta_-^{(k)})^{\alpha_3-1}\end{aligned}\tag{3.4}$$

and

$$\begin{aligned}f(\theta_+^{(k)}|\theta_-^{(k)},\mathbf{x},\mathbf{t}^{(k)}) &\propto \left(\theta_+^{(k)}\right)^{\sum_{i=1}^N(1-t_i^{(k)})x_i+\alpha_2-1}\\ &\times\left(1-\theta_+^{(k)}\right)^{\sum_{i=1}^N(1-t_i^{(k)})(n_i-x_i)}\left(1-\theta_+^{(k)}-\theta_-^{(k)}\right)^{\alpha_3-1}.\end{aligned}\tag{3.5}$$

Sampling from the posterior distribution using priors allowing for conditional dependence requires the Metropolis-Hastings algorithm, and it is likely advisable to use beta densities as the proposal distributions for the misclassification parameters in the Metropolis-Hastings step of the sampler. This is a reasonable approach because we note in Equations 3.4, the full conditional posteriors asymptotically approach beta densities. Furthermore, asymptotically the portion of the posterior related to $(1 - \theta_- - \theta_+)$ simply serves as a constraint on the domain of the parameters. If we set α_3 with a prior value of 1, then this is true for all sample sizes. As a practical matter, we desire a classification system that yields reasonably small misclassification probabilities $(\theta_- + \theta_+ \ll 1)$. Absent this, we will almost certainly fail to produce consistent posterior estimates regardless of the constraints we place on the domain.

3.4.2 Model-Based Approaches for Misclassification Probabilities

One approach of modeling the increased variability due to interindividual factors is to account for the differences in measurements using a mixed effects logistic regression model. For example, we can estimate how misclassification probabilities vary based on the age and/or gender of the respondent, and we can estimate the random variation introduced from random effects such as geographical variation or unobserved individual-level factors. The use of latent class models has often been employed for identification of binary probabilities in the presence of misclassification (Espeland and Hui, 1987). In the stated situation of modeling sensitivities and specificities, a frequentist latent class approach was introduced by Qu et al. (1996) that shares some similarities to our Bayesian approach.

Let us define the two matrices $\mathbf{Z}_-$ and $\mathbf{Z}_+$, where $\mathbf{Z}_-$ is an $n \times p_1$ matrix where column vector $\mathbf{z_1}$ is a column of ones and the remaining columns represent $p_1 - 1$ vectors containing the observed fixed effects information for the FNR, θ_-. Similarly, $\mathbf{Z}_+$ is an $n \times q_1$ matrix with an intercept column followed by $q_1 - 1$ vectors containing all information for the FPR, θ_+. Next, we define $\mathbf{W}_-$ and $\mathbf{W}_+$ as the $n \times p_2$ and $n \times q_2$ design matrices for the random effects for vectors $\boldsymbol{\theta}_-$ and $\boldsymbol{\theta}_+$, respectively. In such a situation we use the mixed models

$$logit(\boldsymbol{\theta}_-) = \mathbf{Z}_-\boldsymbol{\gamma}_{\boldsymbol{-}} + \mathbf{W}_-\boldsymbol{\lambda}_{\boldsymbol{-}} \tag{3.6}$$

and

$$logit(\boldsymbol{\theta}_+) = \mathbf{Z}_+\boldsymbol{\gamma}_+ + \mathbf{W}_+\boldsymbol{\lambda}_+ \tag{3.7}$$

where $\boldsymbol{\gamma}_-$ and $\boldsymbol{\gamma}_+$ are parameter vectors associated with the fixed effects $\mathbf{Z}_-$ and $\mathbf{Z}_+$, and $\boldsymbol{\lambda}_-$ and $\boldsymbol{\lambda}_+$ explain the variability introduced via $\mathbf{W}_-$ and $\mathbf{W}_+$, respectively, to the response probabilities. Often we assume normal distributions on the random effects $\boldsymbol{\lambda}_- \sim N(\mathbf{0}, \boldsymbol{\Sigma}_-)$ and $\boldsymbol{\lambda}_+ \sim N(\mathbf{0}, \boldsymbol{\Sigma}_+)$.

The assumption that $\mathbf{Z}_- = \mathbf{Z}_+$ is quite a temptation given that in most studies, all participants will typically have the same covariate information collected. However, this notation is general enough to allow that different factors may influence the sensitivity and specificity differently. For example, we may wish to maximize the posterior predictive ability of the models by selecting the most parsimonious set of covariates for each of θ_- and θ_+. This would almost certainly result in a different set of variables in $\mathbf{Z}_-$ and $\mathbf{Z}_+$ and likewise for $\mathbf{W}_-$ and $\mathbf{W}_+$.

In the event that only response data was recorded (x_i out of n_i repeated tests for N individuals) and there is still a suspicion that latent interindividual factors that affect the FNR and FPR exist within the data, we turn to a model-based approach to detect this additional variation. Obviously no fixed effects have been observed except in the extremely unusual case that individuals are to be considered fixed. Specifically, suppose the true misclassification probabilities are generated from the models

$$logit(\theta_-) = logit(Pr(Y = 0|T = 1)) = \gamma_- + \epsilon_{-,i} \tag{3.8}$$

$$logit(\theta_+) = logit(Pr(Y = 1|T = 0)) = \gamma_+ + \epsilon_{+,i} \tag{3.9}$$

where each $\epsilon_{-,i} \sim N(0, \sigma_-^2)$ and $\epsilon_{+,i} \sim N(0, \sigma_+^2)$. In this situation, each individual is generated from the same population sharing common misclassification probabilities $\theta_- = logit^{-1}(\gamma_-)$ and $\theta_+ = logit^{-1}(\gamma_+)$, but within each individual the probability of misclassification varies based upon ϵ_- and ϵ_+. The extra "noise" or overdispersion results in data for which extreme cases are observed more often than expected (likewise for underdispersion). For example, if each individual is observed 10 times with misclassification probabilities $\theta_- = 0.1$ and $\theta_+ = 0.1$, it would be highly unlikely to observe a substantial number of individuals with $X_i = 4$, 5, or 6. Likewise, approximately 35% of true positives and true negatives would have $X_i = 10$ or 0, respectively, according to basic binomial probabilities. When the data deviates from this expected behavior substantially, there is likely unaccounted variability in the model, which can be estimated via a random effect.

3.5 Hierarchical Model-Based Approaches for Estimating Misclassification Probabilities

In biostatistics, interindividual variability rarely allows for simplistic assumptions of homogeneous misclassification errors to be satisfied. Consider the situation in which the probability of misclassification may depend on some factor or factors that vary among individuals, such as age, percent body fat, or the presence of a comorbid condition such as diabetes. It may be that the presence or combinations of such factors may make a biomarker more difficult to detect, and thus the probability of misclassification is a function of the set of relevant conditions. Hierarchical models may help facilitate the additional random variability in the event of nonhomogeneous misclassification rates.

3.5.1 The Hierarchical Dirichlet Prior Model

In some cases it is of interest to hierarchically estimate the parameters of the misclassification probabilities θ_-, θ_+, and the related quantity $1 - \theta_- + \theta_+$. For each α_i $(h = 1, 2, 3)$ from (3.3), we can place a *gamma*$(\phi_{1,h}, \phi_{2,h})$ prior distribution recalling that $E[\alpha_i] = \phi_{1,h}/\phi_{2,h}$ and $V[\alpha_i] = \phi_{1,h}/\phi_{2,h}^2$. Each set of $\phi_{1,h}$ and $\phi_{2,h}$ can be set to reflect our prior knowledge of the parameters of the Dirichlet distribution for θ_- and θ_+. Furthermore, if necessary we can utilize diffuse prior distributions to represent our lack of knowledge, such as setting $\phi_{1,h} \geq 1$ and $\phi_{2,h} \leq 1$.

In such a situation, we allow for the ith individual to have his or her own probabilities of misclassification, $(\theta_{-,i}, \theta_{+,i})$, where now we modify (3.3) as follows:

$$p(\theta_{-,i}, \theta_{+,i}) = \frac{1}{B(\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})} \theta_{-,i}^{\alpha_{1,i}-1} \theta_{+,i}^{\alpha_{2,i}-1} (1 - \theta_{-,i} - \theta_{+,i})^{\alpha_{3,i}-1} \tag{3.10}$$

where each $\alpha_{h,i}$ $(h = 1, 2, 3)$ is an observation from

$$p(\alpha_{h,i}|\phi_{1,h}, \phi_{2,h}) = \frac{\phi_{2,h}^{\phi_{1,h}}}{\Gamma(\phi_{1,h})} \alpha_{h,i}^{\phi_{1,h}-1} exp(-\phi_{2,h}\alpha_{h,i}). \tag{3.11}$$

The variability among the parameters $(\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})$ explain the interindividual variation in misclassification probabilities.

The prior distribution on each $\alpha_{h,i}$ in equation (3.11) is completely governed by the values of the hyperparameters $(\phi_{1,h}, \phi_{2,h})$. Specifically, because the variability of each $\alpha_{h,i}$ is strongly governed by $\phi_{2,h}$, especially its value relative to $\phi_{1,h}$, a diffuse prior distribution will have small values for $\phi_{2,h}$. The mean of the prior distribution is less relevant than the variance, as small variance priors have posterior distributions that favor the prior mean and large variance priors yield posteriors that favor the likelihood "mean".

An important subtlety exists if the researcher has prior information. Suppose a pilot study uses a group of truly exposed or diseased individuals $(T = 1)$ to gain information on the false negative rate (θ_-) of the detection system, where $E[\alpha_1/(\alpha_1 + \alpha_2 + \alpha_3)] = \theta_-$. Nothing further is known about the false positive rate (θ_+). To place a small prior variance on the α_1 prior distribution and noninformative distributions on the α_2 and α_3 ultimately yields a posterior distribution with little contribution from the prior information. The reason is that placing a joint prior distribution on (θ_-, θ_+) requires joint knowledge of the two parameters. Thus, any method for eliciting prior information for a single misclassification parameter would necessitate the use of independent distributions such as the beta.

For example, consider the situation in which from our pilot study, 35 of 40 diseased individuals were properly identified as diseased with five false negatives.

We wish to construct a distribution that reflects this information for θ_- using our Dirichlet prior that further reflects a lack of information regarding θ_+. If we assign α_1 a prior distribution with mean 5 (say, the conjugate $gamma(5, 1)$), α_2 a prior distribution with mean 1,(i.e. a $gamma(1, 1)$), and α_3 a prior distribution with mean 39 (i.e. a $gamma(39, 1)$), we induce prior mean $E[\theta_-] = 0.125$, which reflects our prior knowledge with prior standard deviation $\sqrt{(5 \times 35)/(40^2 \times 41)} = 0.052$. However, this also induces prior mean $E[\theta_+] = 0.025$ with prior standard deviation $\sqrt{(1 \times 39)/(40^2 \times 41)} = 0.024$, which has less variability than the parameter for which we have information. Even with a modest amount of prior information on a single parameter as in this case, we have induced a prior distribution on θ_+ that cannot accurately reflect our lack of prior information. The Dirichlet distribution necessarily contains information about both parameters due to the linking quantity $(1 - \theta_- - \theta_+)$, and consequently we require independent priors distributions if we do not have information on both parameters.

3.6 Identifiability

Previous literature on these models from the frequentist realm make clear that the model satisfies identifiability only when 1) $\theta_- + \theta_+ \leq 1$ and 2) the number of repeated measurements of the ith individual is such that $n_i \geq 3$. When θ_- and θ_+ are small, the first limitation is negligible. Either the beta or Dirichlet prior models are sufficient for estimation for small misclassification probabilities. The second limitation is quite easily overcome in the Bayesian realm under certain conditions. Specifically, as long as there exists sufficient information to be able to estimate θ_-,θ_+, and τ, then a subset of the sample can contain individuals with fewer than 3 repeated observations. We must assume that individuals are randomly assigned to the number of repeated tests they receive such that the number of tests (n) is independent of T.

This can be accomplished numerous ways, depending on whether the primary focus is on estimation of τ or θ_- and θ_+. If the misclassification probabilities are of primary interest, a large number of repeated observations relative to the size of the misclassification probability is desirable. If knowledge of the latent τ is of interest, then a larger number of individuals with fewer repeated tests could be gathered. Perhaps the worst case scenario, but still better than assuming the misclassified values to be known, is to use informative prior distributions on the misclassification probabilities based on information gained from previous studies.

In this situation, especially with only 1 or two repeated tests, we may still gain some information for θ_- and θ_+; however, depending on the size of θ_- and θ_+ we can gain substantial information for τ.

3.7 Estimation

One of the primary problems with this approach, especially in detecting whether a random effect exists among individuals, is how many observations are required such that the model benefits from adding a random effect. This problem can be partitioned into two aspects: first, it is of interest to determine how many individuals, N, are required for efficient estimation of the parameters, and secondly, how many repeated observations nested within each individual, n_i, are required for efficient estimation of the random effect. The following simulation explores the model's ability to estimate a random effect under varying levels of N and n_i.

We generated the data set assuming a prevalence $\tau = 0.55$ and misclassification probabilities $\theta_- = 0.14$, and $\theta_+ = 0.05$. For estimation, we are using models 3.8 and 3.9, and thus we define $\gamma_- = logit^{-1}(0.14) \approx -1.815$ and $\gamma_- = logit^{-1}(0.05) \approx -2.944$. We used sample sizes $N = 15, 45$, and 100, and for each sample size we obtained $n_i = 5, 9$, and 12 independent observations from each simulated participant. For each N/n combination, we generated one data set with no random effect and one data set where we added normally distributed random effects to the misclassification parameters

$$logit(\theta_{-,i}) = logit(\theta_{-}) + \epsilon_{-,i},$$

and

$$logit(\theta_{+,i}) = logit(\theta_{+}) + \epsilon_{+,i}$$

where $\epsilon_{-,i} \sim N(0, \sigma_{-}^2)$ where $\sigma_{-} = 0.15$ and $\epsilon_{+,i} \sim N(0, \sigma_{+}^2)$where $\sigma_{+} = 0.25$. We generated and analyzed 200 simulated data sets. Each data set was analyzed twice. The first analysis estimates the model parameters ignoring any random effect, and the second analysis accounts for random effects in the misclassification probability models. For each parameter we present the posterior mean, standard deviation, and 95% credible interval coverage.

For all models, we assume that our prior information for τ can be explained by a $beta(1,1)$ distribution, and that γ_{-} and γ_{+} have $N(0,10)$ prior distributions. For models in which we account for random effects ϵ_{-} and ϵ_{+}, we assume σ_{-} and σ_{+} are distributed $uniform(0,20)$.

3.7.1 Simulation Results

3.7.1.1 *Datasets generated without a random effect.* For each simulated data set, we set initial values for each of three chains. We ran a $5,000$ iteration burn-in followed by a $10,000$ observation sample thinning every third observation. During the sample we monitored τ, γ_{-}, γ_{+}, and in certain cases σ_{-} and σ_{+}.

The first three tables present posterior estimates from data generated without random effects. We compare estimation methods in which we account for a random effect versus estimation where we ignore any potential random effect to elucidate the potential bias introduced by attempting to estimate an effect that may or may not be present. More importantly, given our parameter values, at what points with respect to N and n do we obtain reasonable estimates of the parameters. In Table 3.1, unsurprisingly a sample of 15 individuals yields relatively large posterior estimates of the parameter standard deviations for τ, although the model that includes estimation of a random effect produced substantially larger posterior standard deviations. Furthermore, there is little added benefit to observing repeated observations with regard to τ. It should be noted, the prior mean for τ, which was assigned a $beta(1, 1)$ prior, is 0.5, which explains why the clearly biased and imprecise estimated value of τ from the model that includes random effects could be as close to 0.55 as it is.

The estimated posterior means for τ when estimated without random effect consistently approach the parameter value 0.55 at all combinations of overall sample size N and intra-individual repeated tests n. When the data is estimated where we naïvely include a random effect in the misclassification models, the posterior mean of τ tends to be biased toward the prior mean (0.50), a trend that continues well into the largest sample size of the simulation, especially for small n. Similarly, the posterior mean estimates for θ_- and θ_+ are extremely biased for $N = 15$ when the random effect is included in the model, and this improves only with substantial increases in either the overall sample size or the number of repeated tests.

There is a marked improvement in precision of the posterior distributions for θ_- and θ_+ when measured without a random effect as n grows larger. These improvements are not seen in the models in which a random effect is included, as these models perform extremely poorly under such small sample sizes, especially with a diffuse prior distribution on σ_- and σ_+. Because the data is not generated with a random effect, we can directly observe the instability introduced by superfluous parameters. The prior parameter values for the random effects are centered at mean 10, although the posterior mean should be zero. The proximity of the posterior mean to either 0 or 10 when the random effect has been included in the model is a useful means to determine the relative contribution of the likelihood versus the prior to the posterior value.

As N grows larger, obviously the standard deviations for τ grow smaller, but it also becomes evident that increasing n has an effect of reducing the standard deviations in Tables 3.2 and 3.3, even when the models are estimated with a random effect. The standard deviations for τ measured with and without random effects are nearly the same when $N = 45$ and $n = 12$, and a similar effect occurs for $N = 100$ with $n = 9$ and 12.

The estimation of the posterior mean for σ_- and σ_+ was only attempted for half of the simulations, and given that these data were generated without a random effect, it is evident that we would prefer to observe the model yield posterior estimates for these parameters as close to zero as possible. However, because of the diffuse prior distributions as well as the relatively large means we used for these parameters, this simulation gives us an idea of how much information in terms of sample size we require to overcome the bias introduced via the added variability. The resounding result from this simulation is that under the current parameterization, if there is doubt as to whether a random effect exists, the proposed Bayesian approach achieves posterior mean estimates closer to the generating values when the random effect is omitted.

3.7.1.2 *Datasets generated with a random effect.* Generating datasets that include random effects will inherently increase the variability of the posterior estimates. The need for a simulation is to compare estimates from such datasets in which we capture the additional random variability versus estimates in which we ignore additional random variability. We can directly compare the performance of the two methods and at varying sample sizes to gauge the effect of varying quantities of data on estimates. We observe the results in Tables 3.4, 3.5, and 3.6.

Table 3.1. Simulated datasets for N=15 generated without a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.499	0.544	0.505	0.545	0.506	0.532
StDev	0.27	0.124	0.26	0.115	0.247	0.115
95% CI Coverage	1	0.976	1	0.97	0.998	0.95
γ_- (= -1.815)	0.124	-1.96	0.263	-1.92	0.336	-1.859
StDev	2.52	0.653	2.400	0.368	2.32	0.315
95% CI Coverage	1	0.948	1	0.932	1	0.948
γ_+ (= -2.944)	-0.153	-3.2	-0.391	-3.19	-0.601	-3.11
StDev	2.53	1.14	2.46	0.716	2.4	0.593
95% CI Coverage	1	0.966	1	0.956	0.992	0.936
$\sigma_-(=0)$	7.65	—	6.25	—	5.19	—
StDev	5.2	—	4.9	—	4.47	—
95% CI Coverage	—	—	—	—	—	—
$\sigma_+(=0)$	7.61	—	5.95	—	4.811	—
StDev	5.2	—	4.78	—	4.27	—
95% CI Coverage	—	—	—	—	—	—

Table 3.2. Simulated datasets for N=45 generated without a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.5	0.549	0.516	0.548	0.528	0.544
StDev	0.239	0.073	0.135	0.0714	0.0873	0.0714
95% CI Coverage	1	0.964	0.99	0.946	0.986	0.954
γ_- (= -1.815)	0.208	-1.87	-1.13	-1.84	-1.76	-1.82
StDev	2.2	0.294	1.07	0.201	0.385	0.172
95% CI Coverage	0.996	0.946	0.984	0.956	0.974	0.952
γ_+ (= -2.944)	-0.722	-3.14	-2.37	-3	-3.06	-3
StDev	2.34	0.598	1.46	0.358	0.725	0.304
95% CI Coverage	0.994	0.952	0.984	0.96	0.984	0.946
$\sigma_-(=0)$	6.25	—	2.699	—	1.34	—
StDev	4.7	—	2.43	—	1.2	—
95% CI Coverage	—	—	—	—	—	—
$\sigma_+(=0)$	5.52	—	1.44	—	0.479	—
StDev	4.35	—	1.47	—	0.436	—
95% CI Coverage	—	—	—	—	—	—

Table 3.3. Simulated datasets for N=100 generated without a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.488	0.549	0.543	0.551	0.549	0.55
StDev	0.167	0.0496	0.0556	0.0487	0.0494	0.0487
95% CI Coverage	0.996	0.958	0.958	0.93	0.956	0.948
γ_- (= -1.815)	-0.737	-1.84	-1.88	-1.83	-1.86	-1.82
StDev	1.42	0.189	0.168	0.131	0.129	0.112
95% CI Coverage	0.992	0.954	0.952	0.95	0.934	0.936
γ_+ (= -2.944)	-2.09	-3.03	-3.18	-2.96	-3.12	-2.96
StDev	1.85	0.37	0.419	0.238	0.278	0.202
95% CI Coverage	0.992	0.938	0.932	0.936	0.934	0.962
$\sigma_-(= 0)$	4.68	—	0.918	—	0.57	—
StDev	3.51	—	0.709	—	0.372	—
95% CI Coverage	—	—	—	—	—	—
$\sigma_+(= 0)$	2.57	—	0.347	—	0.284	—
StDev	2.3	—	0.22	—	0.166	—
95% CI Coverage	—	—	—	—	—	—

For sample size $N = 15$ in Table 3.4, we observe similar results as when the data were generated without an individual-level random effect observed in Table 3.1. The posterior estimate for τ scarcely changes from the prior mean, and in general, estimation of the posterior for all parameters is more effective if the random effect is ignored for small sample sizes, regardless of the size of n. The same holds true for the estimation of the posterior probabilities in Table 3.5 at all levels of n except potentially at the highest value, $n = 12$. Thus, when the sample size is moderately large, a large number of repeated tests makes the estimates for the random effects model comparable to the model estimated without random effects. However, we clearly observe that the estimated posterior means for σ_- and σ_+ are still larger than we might hope, although the 95% credible interval coverage improves considerably as n increases.

At the largest sample size in Table 3.6, the posterior means are moderately comparable across the two methods. There is even potential evidence that although the data were generated with an intra-individual random effect, the posterior means measured without the random effect are no worse than the estimation with the random effect. However, the posterior parameter estimates and 95% credible interval coverage also produce estimates that are moderately close to the generating value when $n = 9$ and 12. Therefore we conclude that estimation of random effects in these types of models require moderately large sample sizes (represented here by N) and a moderately large number of repeated tests within each individual. It is fairly evident that 5 repeated tests is too few given the parameter values we have used, and tests on the order of 8 to 10 are required for moderately useful posterior estimates.

It is noteworthy however that this simulation used moderately small values for σ_- and σ_+, and moreover the prior distributions we used were extremely diffuse and not informative of the values we were trying to estimate. The intention of using

Table 3.4. Simulated datasets for N=15 generated with a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.501	0.532	0.504	0.544	0.505	0.539
StDev	0.271	0.124	0.263	0.115	0.246	0.114
95% CI Coverage	1	0.978	1	0.956	1	0.954
γ_- (= -1.815)	0.233	-1.93	0.279	-1.88	0.25	-1.85
StDev	2.51	0.659	2.37	0.365	2.3	0.307
95% CI Coverage	1	0.952	1	0.956	1	0.93
γ_+ (= -2.944)	-0.243	-3.18	-0.385	-3.1	-0.631	-3.14
StDev	2.53	1.13	2.43	0.696	2.4	0.603
95% CI Coverage	0.998	0.968	0.998	0.936	0.992	0.923
σ_-(= 0.15)	7.63	—	6.24	—	5.23	—
StDev	5.14	—	4.88	—	4.44	—
95% CI Coverage	0.118	—	0.184	—	0.402	—
σ_+(= 0.25)	7.57	—	5.99	—	4.66	—
StDev	5.13	—	4.76	—	4.17	—
95% CI Coverage	0.234	—	0.36	—	0.675	—

Table 3.5. Simulated datasets for N=45 generated with a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.501	0.539	0.515	0.552	0.529	0.546
StDev	0.245	0.0731	0.14	0.0714	0.0945	0.0713
95% CI Coverage	1	0.956	0.994	0.952	0.974	0.94
γ_- (= -1.815)	0.327	-1.82	-1.09	-1.83	-1.68	-1.82
StDev	2.2	0.294	1.14	0.199	0.47	0.171
95% CI Coverage	0.998	0.956	0.992	0.96	0.976	0.954
γ_+ (= -2.944)	-0.691	-3.08	-2.26	-2.97	-2.92	-2.95
StDev	2.32	0.579	1.49	0.356	0.811	0.3
95% CI Coverage	0.998	0.938	0.99	0.952	0.986	0.948
σ_-(= 0.15)	6.17	—	2.7	—	1.47	—
StDev	4.65	—	2.44	—	1.28	—
95% CI Coverage	0.104	—	0.522	—	0.732	—
σ_+(= 0.25)	5.59	—	1.55	—	0.617	—
StDev	4.43	—	1.62	—	0.561	—
95% CI Coverage	0.338	—	0.896	—	0.986	—

Table 3.6. Simulated datasets for N=100 generated with a random effect

	Posterior Estimates					
	5 Repeated Binary Tests		9 Repeated Binary Tests		12 Repeated Binary Tests	
Parameter	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects	With Random Effects	Without Random Effects
τ (=0.55)	0.49	0.549	0.538	0.548	0.548	0.55
StDev	0.17	0.0498	0.0562	0.0488	0.05	0.0487
95% CI Coverage	1	0.954	0.978	0.952	0.952	0.946
γ_- (= -1.815)	-0.714	-1.83	-1.88	-1.82	-1.88	-1.83
StDev	1.43	0.19	0.172	0.131	0.132	0.112
95% CI Coverage	0.996	0.95	0.964	0.954	0.95	0.95
γ_+ (= -2.944)	-2.04	-2.98	-3.18	-2.96	-3.11	-2.95
StDev	1.84	0.362	0.412	0.237	0.285	0.201
95% CI Coverage	0.978	0.954	0.954	0.946	0.942	0.966
σ_-(= 0.15)	4.6	—	0.953	—	0.603	—
StDev	3.44	—	0.726	—	0.402	—
95% CI Coverage	0.154	—	0.726	—	0.824	—
σ_+(= 0.25)	2.58	—	0.36	—	0.307	—
StDev	2.29	—	0.223	—	0.174	—
95% CI Coverage	0.702	—	0.988	—	0.994	—

such small parameter values was that on the logit scale the added variability would likely be detectable without adding excessive noise to the data. However, including the possibility of random effects yielded noisy estimates. A more flexible or even informative prior distribution may have improved the performance of the models, although this would require a diffuse gamma prior distribution on σ_- and σ_+, which have documented difficulties when implemented in MCMC algorithms.

3.8 Applications

We demonstrate the methods using two datasets. The first was published by Boyles (2001) and serves as a simple demonstration of the method from the quality control literature. The second data set was published by Fujisawa and Izumi (2000) originally and demonstrates the advantages of the Bayesian approach for parameter estimation over previously published methods. For all simulations, we use WinBUGS v1.4.3 (Lunn et al., 2000).

3.8.1 Quality Control Data Example

The first data set features 9 repeated observations among 38 independent units from inkjet print-sample inspection data. The data are presented on Table 3.7 and feature the frequencies of "Pass" test results.

Table 3.7. Repeated Binary Test Frequencies for 38 Subjects

# of tests					$n_i = 9$					
x_i	0	1	2	3	4	5	6	7	8	9
Frequency	10	2	2	1	2	0	0	3	2	16

We estimate the probability that any randomly selected item, T_i, is acceptable, $\tau = pr(T_i = 1)$, as well as the false positive rate and false negative rate, θ_+ and θ_-, respectively. We use the posterior distribution described in 3.2 and subsequent Gibbs sampler, where we assume $beta(1, 1)$ prior distributions on each of τ, θ_+, and θ_-. We set initial values for each of 3 chains, run a 1000 iteration burn-in, and collect samples of 1000 observations from the posterior. The posterior results are presented in Table 3.8.

Table 3.8. Assumes beta(1,1) priors on all parameters

	Posterior Mean	Posterior St Dev	95% Credible Interval
τ	0.55	0.078	(0.39, 0.70)
θ_-	0.047	0.016	(0.022, 0.082)
θ_+	0.12	0.026	(0.070, 0.17)

Suppose the sampled inkjet cartridges were manufactured on different days or at different production facilities. In such a situation, we might assume that there exist some inherent heterogeneity among the individual cartridges that potentially cannot be measured directly other than in the added variability in the response. For example, looking at the data in Table 3.7 and the results in Table 3.8, we might be suspicious of additional components of variability when we observe that $\sum_{i=1}^{N} I(x_i = 3) < \sum_{i=1}^{N} I(x_i = 4)$ and $\sum_{i=1}^{N} I(x_i = 7) > \sum_{i=1}^{N} I(x_i = 8)$ where $I(\cdot)$ is the indicator function. This potentially suggests additional variability in θ_- and θ_+ that may require the use of hierarchical Bayesian modeling. Furthermore, the preceding simulations show that with a moderately large sample size and moderately large number of repeated observations, it is possible to detect inter-individual random variability that exists

To allow for a hierarchical model, we assume that for the ith individual $\theta_{-,i}$ and $\theta_{+,i}$ are sampled from a $Dirichlet(\alpha_1, \alpha_2, \alpha_3)$ distribution. Lacking further prior information, we place $gamma(4, 1)$ hyperprior distributions upon each $\alpha_j (j = 1, 2, 3)$. This allows for estimation of a unique τ, 38 realizations each of θ_- and θ_+, and full posterior estimates of each α_j. We denote $\widetilde{\theta_-}$ as the set of 38 estimated posterior

Table 3.9. Posterior estimates of hierarchical Dirichlet model

	Posterior Mean	Posterior St Dev	95% Credible Interval
τ	0.56	0.081	(0.40, 0.72)
α_1	0.72	0.30	(0.29, 1.4)
α_2	1.2	0.49	(0.45, 2.3)
α_3	7.5	2.5	(3.4, 13)
$\widetilde{\theta_-}$	0.072	0.034	(0.038, 0.082)
$\widetilde{\theta_+}$	0.12	0.026	(0.070, 0.17)

false negative rates, and $\widetilde{\theta_+}$ as the set of 38 estimated posterior false positive rates. We observe the results in Table 3.9.

Obviously we have changed the prior information placed on the misclassification parameters in both distribution and in quantity. The larger the value of the first parameter in the gamma hyperprior distribution, the more "informative" the prior distribution for each of the α_j elements. Thus, in our example where each α_j is assumed to be equally unknown, we have a posterior that is slightly biased toward equal values of α_1, α_2, and α_3. This is an unfortunate consequence of using a gamma hyperprior distribution on the elements of the Dirichlet distribution in this case. The relatively small posterior value for α_1 necessitated a slightly larger prior distribution mean to avoid a gamma distribution that samples too close to the lower bound of zero.

3.8.2 Biostatistical Application

We next employ our model on a data set of tests for the MNSs antigen group within atomic bomb survivors and their children. The data are presented in Table 3.10. We observe that each individual was assessed n = 2, 3, or 4 times, and because blood antigen status should not vary by time, the testing method clearly contains misclassification. However, there is evidence that intraindividual factors may affect the ability to detect these antigens, especially participant age and health status.

Table 3.10. Test Quality Control Data from Fujisawa & Izumi, 2000

		Frequencies of Positive Tests given n_i											
	# of tests	$n_i = 2$			$n_i = 3$				$n_i = 4$				
Antigen	City	0	1	2	0	1	2	3	0	1	2	3	4
M	Hiroshima	419	8	1918	77	4	1	279	4	1	0	0	29
	Nagasaki	257	13	958	26	2	1	127	3	0	0	0	13
N	Hiroshima	714	23	1587	117	5	10	225	13	0	0	2	19
	Nagasaki	324	70	799	40	3	27	85	4	1	0	4	7
S	Hiroshima	1823	29	208	269	1	10	33	24	1	0	2	1
	Nagasaki	868	52	43	83	1	7	4	8	0	0	0	0
s	Hiroshima	19	1	2316	9	0	0	349	0	0	0	1	33
	Nagasaki	5	5	1065	1	1	3	133	0	0	0	0	15

The majority of tested individuals were tested only twice, although we require $n \geq 3$ to maintain identifiability. As discussed previously, the Bayesian approach to parameter estimation can easily overcomes this in two ways. First, if we possess prior information on the parameters, we can represent prior knowledge on the misclassification parameters with a distribution that will allow for identifiable estimation of τ for when $n < 3$; likewise, we may estimate the misclassification parameters with known prior information on τ. Secondly, if we possess a moderately large number of individuals with $n \geq 3$, then this information can be incorporated into the model to gain additional knowledge of τ for individuals with fewer than three repeated binary tests.

Our data set satisfies the latter condition. We require the assumption that the number of repeated binary tests are randomly assigned to individuals; thus, we assume the probability of misclassification is independent of n. Otherwise, we make few prior assumptions about the parameters by assigning $beta(1, 1)$ prior distributions to each of τ, θ_{-}, and θ_{+}. Next, we sample according to the scheme described for the posterior in 3.1. During the sampling scheme, we obtain estimates of θ_{-} and θ_{+} from the individuals with $n \geq 3$ and allow these estimates to apply guide the additional information we obtain from τ where $n = 2$.

Table 3.11. Posterior estimates for model with beta(1,1) priors

Antigen	City	Parameter	Mean	St Dev	0.025%	0.975%
M	Hiroshima	τ	0.813	0.00744	0.799	0.828
		θ_-	0.000817	0.000547	0.0000831	0.00215
		θ_+	0.0110	0.00353	0.00495	0.0187
	Nagasaki	τ	0.788	0.0111	0.766	0.809
		θ_-	0.00292	0.00175	0.000334	0.00693
		θ_+	0.0181	0.00747	0.00530	0.0342
N	Hiroshima	τ	0.685	0.00903	0.667	0.702
		θ_-	0.00734	0.00162	0.00440	0.0108
		θ_+	0.00706	0.00265	0.00269	0.0129
	Nagasaki	τ	0.723	0.0128	0.698	0.748
		θ_-	0.0475	0.00559	0.0369	0.0589
		θ_+	0.0148	0.00795	0.00269	0.0331
S	Hiroshima	τ	0.117	0.00680	0.104	0.130
		θ_-	0.0650	0.0123	0.0423	0.0903
		θ_+	0.00169	0.00102	0.000232	0.00413
	Nagasaki	τ	0.103	0.0144	0.0740	0.131
		θ_-	0.320	0.0529	0.210	0.419
		θ_+	0.00494	0.00411	0.000161	0.0154
s	Hiroshima	τ	0.989	0.00197	0.985	0.993
		θ_-	0.000461	0.000294	0.0000692	0.00118
		θ_+	0.0199	0.0189	0.000579	0.0698
	Nagasaki	τ	0.990	0.00452	0.979	0.996
		θ_-	0.00227	0.00130	0.000180	0.00513
		θ_+	0.280	0.152	0.0417	0.600

The results presented in Table 3.11 display the posterior information regarding each parameter for each location and antigen type. As we might expect the prevalence τ for unique antigen types is quite similar across locations. Furthermore, the probabilities of misclassification are uniformly small with very few exceptions. Even in situations in which the misclassification probabilities are greater than 0.10, we must note that the prior means of each of the parameters is 0.5, and that in these cases there is an extreme value for τ, reducing the number of true positives or true negatives even with a large sample size.

The conclusions of the first model give evidence that the antigen type across the two cities have similar prevalences. Furthermore, because the misclassification probabilities are generally small (with a couple of exceptions), we could model these as being generated with a common mean and the deviations from this mean be attributable to a random effect. The model we propose to simplify the estimation process uses the 4 element design vector $\mathbf{z}' = (1, Z_1, Z_2, Z_3)$ where Z_1 is an indicator variable such that $Z_{1,i} = 1$ if the ith individual has antigen N and 0 otherwise, and correspondingly, $Z_{2,i} = 1$ for antigen S and $Z_{3,i} = 1$ for antigen s. Then we define

$$
\begin{aligned}
&logit(\tau_k) = \mathbf{z}'\boldsymbol{\beta}, \\
&logit(\theta_{-,i}) = \gamma_- + \epsilon_{-,i}, \\
&logit(\theta_{+,i}) = \gamma_+ + \epsilon_{+,i}
\end{aligned}
$$

for $k = 1, \ldots, 4$ where τ_k is the posterior prevalence of the given antigen, and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \beta_3)$ where intercept β_0 represents the log-odds of antigen M, and β_1, β_2, and β_3 represent the change in the log-odds corresponding to antigens N, S, and s, respectively. Furthermore, $\theta_{-,i}$ and $\theta_{+,i}$ represent the false negative and false positive rates, respectively, for $i = 1, \ldots, N$. We define γ_- and γ_+ as the log-odds of the common mean false negative and false positive rates, respectively, and $\epsilon_{-,i}$ and $\epsilon_{+,i}$ are the respective random effects, where we assume $\epsilon_{-,i} \sim N(0, \sigma_-^2)$ and $\epsilon_{+,i} \sim N(0, \sigma_+^2)$.

We have observed no prior information, and thus we utilize moderately diffuse prior distributions for each parameter. We set

$$\boldsymbol{\beta} \sim N(\mathbf{0}, 100\mathbf{I}_4),$$

$$\gamma_- \sim N(0, 100),$$

$$\gamma_+ \sim N(0, 100),$$

$$\sigma_- \sim gamma(1, 1),$$

$$\sigma_+ \sim gamma(1, 1),$$

where $\mathbf{I}_4$ is the 4×4 identity matrix.

We initialize the chains, reject the first 1000 burn-in observations, and then collect a sample of 10,000 posterior observations from the WinBUGS sampler, thinning every third observation. The posterior estimates of means and variability terms are presented in Table 3.12.

Table 3.12. Posterior estimates for Hiroshima survivors model-based estimates

Parameter	Mean	St Dev	0.025%	0.975%
β_0	1.42	0.040	1.34	1.50
β_1	-0.58	0.053	-0.68	-0.48
β_2	-3.48	0.072	-3.63	-3.34
β_3	3.29	0.180	2.94	3.65
γ_-	-5.05	0.86	-6.91	-3.50
γ_+	-5.10	0.59	-6.28	-3.95
σ_-	2.53	0.66	1.57	4.10
σ_+	1.17	0.64	0.26	2.76

3.9 Discussion

This is a straightforward extension of the previous work done on pass/fail testing systems with the added benefit of Bayesian inference. The addition of hierarchical modeling in this context opens possibilities for the broader use of these methods in the biostatistical arena by controlling for and estimating the effects of interindividual variability. The assumption that the misclassification probabilities are constant over time could be relaxed in future research to allow for new advances in diagnostic capabilities.

CHAPTER FOUR

Bayesian Approaches to Detecting Dependent Binary Data

4.1 Introduction

Binary misclassification is a common phenomenon in the biological and epidemiological sciences that results in biased statistical inference when adjustment for the misclassification is not performed. Approaches to statistical inference that accounts for misclassification has a lengthy history in the frequentist literature (see Tenenbein (1970), Hochberg (1977), and Espeland and Hui (1987)). In the late twentieth century, the Bayesian approach to statistical inference flourished thanks to developments in the applicability of MCMC methods (Gelfand and Smith, 1990) and the increasing speed and ubiquity of computing power. Some major developments in Bayesian approaches to statistical estimation in the presence of misclassified binary data include Paulino et al. (2003), Joseph et al. (1995), and Gustafson (2004).

Correction for binary misclassification can be an expensive, time-consuming, and often invasive process because typically we require at least two observations by which we attempt to quantify the test probabilities of correct classification, the sensitivity and specificity. Furthermore, it requires planning in the design process to allocate time, money, etc. to collect sufficient data by which we validate our fallible classifier. When a perfect classifier, or gold standard test, exists, we may choose to test all study participants using our fallible classifier and utilize the gold standard test on a subsample of the total study population (Tenenbein, 1970). This method, known as double sampling, is generally utilized when the gold standard is exceedingly expensive or introduces unacceptable risk to the larger study population.

Many situations require measurement in which a gold standard test does not exist or is not safely available for use. Such tests rely on two or more imperfect measurement systems without the convenience of comparison with a perfect classifier. For example, Fujisawa and Izumi (2000) utilized a repeated binary diagnostic test to estimate the prevalence of blood-based antigens among a longitudinal study group in Japan. Another tool is the dual test protocol, which relies on two fallible binary classifiers to estimate the latent true classification status. Its use has been documented by Joseph et al. (1995) in a Bayesian context for the serological examination testing for *Strongyloides* infection among Cambodian refugees in Canada. One assumption that was made by Joseph et al. (1995) that was later adopted by Ren and Stone (2007a) was that the tests were conditionally independent, or that the processes by which one test fails to correctly classify an observation has nothing to do with the processes by which the second test misclassifies an individual.

A more general approach developed by Black and Craig (2002) allows for potential dependence between diagnostic test sensitivities and/or specificities. However, they recognize that any such method suffers from a lack of identifiability of model parameters, requiring informative prior assumptions on model parameters to estimate the unknowns. This is unrealistic in many situations as our reliance on a dual test protocol usually implies a lack of information on the test properties.

It is important to have some knowledge of the impact of misspecification on the dependence assumption of a dual test protocol. We attempt to evaluate the impact of ignoring misclassification parameter dependence on posterior parameter estimates. Additionally, we use noninformative prior distributions to properly account for dependent misclassification probabilities to evaluate whether a nonidentifiable Bayesian model can produce reliable posterior estimates in such a situation. The final aim of the chapter is to evaluate statistical methods to detect conditional dependence between binary diagnostic tests in the context of no or little prior data.

In Section 4.2 we formulate the general likelihood and posterior distribution assuming an independent dual test protocol. In Section 4.3 we add dependence parameters between the sensitivities and specificities and formulate the full posterior distribution. In Sections 4.4.1 and 4.4.2 we describe two statistical approaches for detecting conditional dependence in binary diagnostic tests, and in Section 4.5 we evaluate the posterior estimates and performance of the model evaluation methods in detecting dependence in binary tests. We apply the methods to a previously published data set in 4.6.

4.2 Independent Binary Tests

In a simple logistic regression situation with multiple covariates, we assume that the relationship between the binomial response Y and the covariate vector $\mathbf{z}$ of length p can be explained by the relationship

$$logit(\pi) = \mathbf{z}'\boldsymbol{\beta} \tag{4.1}$$

where $\pi = pr(Y = 1|\mathbf{z})$ and vector $\boldsymbol{\beta}' = (\beta_0, \ldots, \beta_p)$ where β_k represents the log-odds contribution of the kth element of vector $\mathbf{z}$. Consistent estimation of π relies on perfectly observed response and covariate values.

4.2.1 Misclassification

Suppose we introduce an additional binary covariate that is measured with error due to a fallible classification system. Because of the error, it is actually impossible to observe its true status, represented by T, forcing us to rely on fallible classifiers that may produce results inconsistent with the truth. One such method is the dual test protocol, which is efficient due to its reliance upon two binary tests. Using such a system allows us to estimate the true exposure status of T, the covariate of interest, where we assume T is distributed $Bernoulli(\tau)$. We denote the two fallible binary classifiers as X_1 and X_2, where we define the sensitivities $S_1 = pr(X_1 = 1|T = 1)$ and $S_2 = pr(X_2 = 1|T = 1)$ and the specificities $C_1 = pr(X_1 = 0|T = 0)$ and $C_2 = pr(X_2 = 0|T = 0)$. If the two tests are independent, the joint distribution of the observed X_1 and X_2 given the latent T is

$$\begin{aligned} f(x_1, x_2|t) = &\left(S_1^{x_1}(1-S_1)^{1-x_1}S_2^{x_2}(1-S_2)^{1-x_2}\right)^t \\ &\times \left((1-C_1)^{x_1}C_1^{1-x_1}(1-C_2)^{x_2}(C_2)^{1-x_2}\right)^{1-t}. \end{aligned} \tag{4.2}$$

To arrive at the full joint distribution of the observed variables X_1, X_2, and the latent T, we employ the distribution of T such that

$$\begin{aligned} f(x_1, x_2, t) &= f(x_1, x_2|t)f(t) \\ &= \left(\tau S_1^{x_1}(1-S_1)^{1-x_1}S_2^{x_2}(1-S_2)^{1-x_2}\right)^t \\ &\times \left((1-\tau)(1-C_1)^{x_1}C_1^{1-x_1}(1-C_2)^{x_2}(C_2)^{1-x_2}\right)^{1-t}. \end{aligned}$$

We can then sum latent T from the statement by noting $f(x_1, x_2) = \sum_{t=0}^{1} f(x_1, x_2, t)$ to obtain a distribution that will ultimately allow us to make inference upon our parameters based solely upon the observed data. Thus, we denote the joint distribution of independent X_1 and X_2 as

$$\begin{aligned} f(x_1, x_2) = &\left(\tau S_1 S_2 + (1-\tau)(1-C_1)(1-C_2)\right)^{I(x_1=1,x_2=1)} \\ &\times \left(\tau S_1(1-S_2) + (1-\tau)(1-C_1)C_2\right)^{I(x_1=1,x_2=0)} \\ &\times \left(\tau(1-S_1)S_2 + (1-\tau)C_1(1-C_2)\right)^{I(x_1=0,x_2=1)} \\ &\times \left(\tau(1-S_1)(1-S_2) + (1-\tau)C_1C_2\right)^{I(x_1=0,x_2=0)}, \end{aligned} \tag{4.3}$$

where $I(\cdot)$ is the indicator function. The joint distribution of N independent observations takes the form of a multinomial distribution.

We note that we can only observe X_1 and X_2 with four possible outcomes: $X_1 = X_2 = 0$; $X_1 = 0, X_2 = 1$; $X_1 = 1, X_2 = 0$; and $X_1 = X_2 = 1$. Given that there are five parameters ($S_1, S_2, C_1, C_2,$ and τ), this model is not identifiable. We can gain additional information if the value of τ varies based on our perfectly observed covariates $\mathbf{z}$ via the logistic regression model-based identity

$$f(t|\mathbf{z}) = pr(T = t|\mathbf{z}) = \frac{exp\left(t(\mathbf{z}'\boldsymbol{\lambda})\right)}{1 + exp(\mathbf{z}'\boldsymbol{\lambda})},$$

where $\boldsymbol{\lambda}$ is the parameter vector that relates $\mathbf{z}$ to τ. If there is substantial separation between the prevalence of τ at different levels of $\mathbf{z}$, that is, for significantly large values of $\boldsymbol{\lambda}$, then identifiability will likely be achieved (Cheng et al., 2007).

We allow the mean of our outcome variable Y to vary based upon $\mathbf{z}$ and T, such that we can rewrite 4.1 as

$$f(y|t, z) = pr(Y = y|\mathbf{z}, t) = \frac{exp\,(y(\mathbf{z}'\boldsymbol{\beta} + \phi t))}{1 + exp(\mathbf{z}'\boldsymbol{\beta} + \phi t)},$$

where ϕ is the coefficient for latent T.

Using these relationships, we develop our likelihood for the ith individual as

$$f(y_i, x_{1,i}, x_{2,i}, \mathbf{z}_i|t_i, S_1, S_2, C_1, C_2, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi) = f(y_i|t_i, \mathbf{z}_i)f(t_i|\mathbf{z}_i)f(x_{1,i}, x_{2,i}|t_i). \quad (4.4)$$

We next define the prior information that we have for our parameters using the general notation

$$f(\boldsymbol{\beta}, \phi, \boldsymbol{\lambda}, S_1, S_2, C_1, C_2)$$

to represent our prior distribution. Assuming a sample of size N observations, where we denote $\mathbf{y}' = (y_1, \ldots, y_N)$, $\mathbf{x}_1 = (x_{1,1}, \ldots, x_{1,N})$, etc., we combine the likelihood and prior distributions in the manner of Ren and Stone (2007a) to obtain the posterior distribution

$$\begin{aligned}
&f(\boldsymbol{\beta}, \phi, \boldsymbol{\lambda}, S_1, S_2, C_1, C_2, \mathbf{t}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \mathbf{Z}) \\
&\quad = f(\boldsymbol{\beta}, \phi, \boldsymbol{\lambda}, S_1, S_2, C_1, C_2) \prod_{i=1}^{N} f(y_i, x_{1,i}, x_{2,i}, \mathbf{z}_i|t_i, S_1, S_2, C_1, C_2, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi) \\
&\quad = f(\boldsymbol{\beta}, \phi, \boldsymbol{\lambda}, S_1, S_2, C_1, C_2) \left\{ \prod_{i=1}^{N} \frac{exp\,(y_i(\mathbf{z}_i'\boldsymbol{\beta} + \phi t_i))}{1 + exp(\mathbf{z}_i'\boldsymbol{\beta} + \phi t_i)} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{N} \left(S_1^{x_{1,i}} (1 - S_1)^{1-x_{1,i}} S_2^{x_{2,i}} (1 - S_2)^{1-x_{2,i}} \right)^{t_i} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{N} \left((1 - C_1)^{x_{1,i}} C_1^{1-x_{1,i}} (1 - C_2)^{x_{2,i}} (C_2)^{1-x_{2,i}} \right)^{1-t_i} \right\} \\
&\quad \times \left\{ \prod_{i=1}^{N} \frac{exp\,(t(\mathbf{z}_i'\boldsymbol{\lambda}))}{1 + exp(\mathbf{z}_i'\boldsymbol{\lambda})} \right\}.
\end{aligned}$$

4.3 Dependent Binary Tests

Suppose that X_1 and X_2 are dependent tests such that $pr(X_1 = 1, X_2 = 1|T = 1) \neq S_1 \times S_2$ and/or $pr(X_1 = 0, X_2 = 0|T = 0) \neq C_1 \times C_2$. Let us define conditional dependence term for the sensitivities $\rho_S = pr(X_1 = 1, X_2 = 1|T = 1) - S_1S_2$ where $\rho_S \in ((S_1 - 1)(1 - S_2), min(S_1, S_2) - S_1S_2)$ Vacek (1985). Likewise, the conditional dependence term for the specificities $\rho_C = pr(X_1 = 0, X_2 = 0|T = 0) - C_1C_2$ is defined on the interval $((C_1-1)(1-C_2), min(C_1, C_2)-C_1C_2)$. To add the dependence terms to the data likelihood, we simply modify the joint density of X_1 and X_2 such that

$$\begin{aligned} f(x_1, x_2|t_i) = {} & (t_i(S_1S_2 + \rho_S) + (1 - t_i)((1 - C_1)(1 - C_2) + \rho_C))^{I(x_1=1,x_2=1)} \\ & \times (t_i(S_1(1 - S_2) - \rho_S) + (1 - t_i)((1 - C_1)C_2 - \rho_C))^{I(x_1=1,x_2=0)} \\ & \times (t_i((1 - S_1)S_2 - \rho_S) + (1 - t_i)(C_1(1 - C_2) - \rho_C))^{I(x_1=0,x_2=1)} \\ & \times (t_i((1 - S_1)(1 - S_2) + \rho_S) + (1 - t_i)(C_1C_2 + \rho_C))^{I(x_1=0,x_2=0)} . \end{aligned} \tag{4.5}$$

It is then straightforward to obtain the marginal density for the dual test protocol as

$$\begin{aligned} f(x_1, x_2) = {} & (\tau(S_1S_2 + \rho_S) + (1 - \tau)((1 - C_1)(1 - C_2) + \rho_C))^{I(x_1=1,x_2=1)} \\ & \times (\tau(S_1(1 - S_2) - \rho_S) + (1 - \tau)((1 - C_1)C_2 - \rho_C))^{I(x_1=1,x_2=0)} \\ & \times (\tau((1 - S_1)S_2 - \rho_S) + (1 - \tau)(C_1(1 - C_2) - \rho_C))^{I(x_1=0,x_2=1)} \\ & \times (\tau((1 - S_1)(1 - S_2) + \rho_S) + (1 - \tau)(C_1C_2 + \rho_C))^{I(x_1=0,x_2=0)} . \end{aligned} \tag{4.6}$$

From our likelihood distribution expressed in 4.4, we note that the inclusion of dependence parameters only affects $f(x_1, x_2|t)$, but the conclusions necessarily have a bearing on the estimation of latent T, and using 4.3 when in fact the tests are dependent subsequently introduces bias into the conclusions. Estimation of π and subsequently S_1, S_2, C_1, and C_2 depend on all other parameters through the MCMC algorithm as demonstrated by Ren and Stone (2007a).

We assume in the preceding model similarly to Dendukuri and Joseph (2001) that the test sensitivities and specificities are homogeneous across values of $\mathbf{z}$. This simplifying assumption should necessarily be scrutinized; however, the loss of the homogeneity assumption of our model eliminates the already marginal identifiability of the data, resulting in the need for the use of informative prior distributions to achieve posterior parameter estimates. We proceed assuming homogeneous misclassification probabilities as we assume no further prior information on our parameters.

4.4 Model Assessment Methods

Because parameter identifiability is such a concern, the introduction of a dependence term or terms will create an additional challenge for estimation from the posterior distributions. On the other hand, to simply ignore the association introduces potential bias into the parameter estimates.

4.4.1 Gibbs Variable Selection

In order to achieve parsimony and develop objective criteria regarding whether to include or exclude dependence parameters we extend a Bayesian variable selection procedure to this context in which model identifiability becomes a major concern. We use the method of Gibbs Variable Selection proposed by Dellaportas et al. (2002) in which we use the binary indicator priors for variable selection in regression (Kuo and Mallick, 1998) to identify the most parsimonious model. We create $k = 2$ indicator variables γ_1 and γ_2 and multiply them by ρ_S and ρ_C, respectively, and allow the MCMC process to identify the model of highest posterior value. We thus modify 4.5 as follows:

$$\begin{aligned} f(x_1, x_2|t_i) = {} & (t_i(S_1S_2 + \rho_S\gamma_1) + (1-t_i)((1-C_1)(1-C_2) + \rho_C\gamma_2))^{I(x_1=1,x_2=1)} \\ & \times (t_i(S_1(1-S_2) - \rho_S\gamma_1) + (1-t_i)((1-C_1)C_2 - \rho_C\gamma_2))^{I(x_1=1,x_2=0)} \\ & \times (t_i((1-S_1)S_2 - \rho_S\gamma_1) + (1-t_i)(C_1(1-C_2) - \rho_C\gamma_2))^{I(x_1=0,x_2=1)} \\ & \times (t_i((1-S_1)(1-S_2) + \rho_S\gamma_1) + (1-t_i)(C_1C_2 + \rho_C\gamma_2))^{I(x_1=0,x_2=0)} . \end{aligned} \tag{4.7}$$

Each γ_j $(j = 1, 2)$ is assigned a $Bernoulli(0.5)$ prior distribution which induces four possible models of equal prior probability:

(1) the independent model, in which $\gamma_1 = \gamma_2 = 0$ and the tests are subsequently considered independent,

(2) the dependent sensitivity model, in which $\gamma_1 = 1$ and $\gamma_2 = 0$,

(3) the dependent specificity model, in which $\gamma_1 = 0$ and $\gamma_2 = 1$, and

(4) the dual dependence model where both $\gamma_1 = \gamma_2 = 1$.

A method for sampling such a model from WinBUGS is explained in Ntzoufras (2002).

4.4.2 Deviance Information Criteria

The Deviance Information Criteria (DIC) is a Bayesian measure for assessing model fit in which incorporates the posterior deviance, or fit of the data given the selected parameters, penalized by the additional complexity due to adding extra parameters as explained by Spiegelhalter et al. (2002) in their seminal paper. This measure gives similar results as Akaike's Information Criteria (Akaike, 1973) under negligible prior information, but its subsequent strength is its ability to produce model assessments under informative or noninformative prior assumptions. We define posterior distribution $p(\theta|y)$ with parameter set θ and data y such that $E[\theta|y] = \bar{\theta}$, which implies $\bar{\theta}$ is simply the set of posterior parameter means. Next, we define $D(\bar{\theta}) = -2log(p(y|\bar{\theta})) + 2log(f(y))$, where for $E[Y] = \mu(\theta)$ then $f(y) = p(y|\mu(\theta) = y)$. Then

$$DIC = D(\bar{\theta}) + 2p_D,$$

where p_D is roughly interpreted as the expected degree of "overfitting" of the posterior estimates $\bar{\theta}$ to the data y. Strictly speaking, for model with focus of inference

Θ, then

$$p_D\{Y, \Theta, \bar{\theta}\} = E_{\theta|y}[-2log(p(y|\theta))] + 2log[p(y|\bar{\theta}(y))]$$

is the proposed effective number of parameters.

Because the DIC increases with p_D, its value should increase with superfluously added parameters and decrease where the fit genuinely improves without overfitting. However, we use it in the context of attempting to determine the fit of an already saturated model in the attempt to determine whether its performance can identify the data generating model in a noninformative prior setting.

4.5 Simulation

We note that previous explorations of this model assuming an independent dual test protocol analyzed a single simulated data set rather than a series of simulated data sets (Ren and Stone (2007a),Ren and Stone (2007b)). Furthermore, when considering the more general approach taken by Black and Craig (2002), we note that a relatively large amount of prior information was used for model evaluation and subsequent estimation of model parameters. Our approach is to utilize the diffuse prior approach of Ren and Stone (2007a) in the attempt to estimate the parameters of various combinations of dependent binary tests in models similar to that of Black and Craig (2002).

In our simulations, we address three primary concerns:

(1) we attempt to quantify the model instability that arises when we include conditional dependence in our analysis when in reality the dual test protocol is independent.

(2) we quantify the bias introduced when the data analysis ignores conditional dependence when in reality the dual test protocol contains conditionally dependent process.

(3) we detect the data-generating model via the use of Gibbs Variable Selection and DIC statistic discussed in sections 4.4.1 and 4.4.2.

In all situations we assume that no prior information exists and thus we use diffuse priors on all parameters. This adds a computational burden as the MCMC chain must attempt to estimate unique parameter values in a near-nonidentifiable model.

For each of four scenarios, we generate 200 datasets of $N = 800$ observations. For the ith observation within each dataset we generate binary response y_i, "true" classification status t_i, fallible dual binary test protocol outcomes $x_{1,i}$ and $x_{2,i}$, and design vector $\mathbf{z}_i' = \begin{pmatrix} 1 & z_{1,i} & z_{2,i} & z_{3,i} \end{pmatrix}$, where for $h = 1, 2, 3$, binary $z_{h,i}$ indicates the ith participant's inclusion in one of four mutually exclusive groups. We further define

$$logit(pr(t_i = 1|\mathbf{z}_i)) = \mathbf{z}_i'\boldsymbol{\lambda} = -0.85 - z_{1,i} + z_{2,i} + 2z_{3,i},$$

$$logit(pr(y_i = 1|t_i, \mathbf{z}_i)) = \mathbf{z}_i'\boldsymbol{\beta} + \phi t_i = -0.7 - 0.3z_{1,i} + 0.8z_{2,i} + 1.6z_{3,i} + 1.4t_i,$$

$$pr(x_{1,i} = 1, x_{2,i} = 1|t_i) = t_i(S_1S_2 + \rho_S) + (1 - t_i)((1 - C_1)(1 - C_2) + \rho_C),$$

$$pr(x_{1,i} = 1, x_{2,i} = 0|t_i) = t_i(S_1(1 - S_2) - \rho_S) + (1 - t_i)((1 - C_1)C_2 - \rho_C),$$

$$pr(x_{1,i} = 0, x_{2,i} = 1|t_i) = t_i((1 - S_1)S_2 - \rho_S) + (1 - t_i)(C_1(1 - C_2) - \rho_C),$$

$$pr(x_{1,i} = 0, x_{2,i} = 0|t_i) = t_i((1 - S_1)(1 - S_2) + \rho_S) + (1 - t_i)(C_1C_2 + \rho_C),$$

where $S_1 = 0.9, S_2 = 0.7, C_1 = 0.75$, and $C_2 = 0.95$ for all simulated data sets. Although the data will be generated with T, we will then assume that T is latent and proceed with posterior estimation with only Y, X_1, X_2, and $\mathbf{z}$.

The four subsections diverge with respect to the quantities ρ_S and ρ_C. In Section 4.5.1 we generate X_1 and X_2 with $\rho_S = \rho_C = 0$, which implies the two tests are conditionally independent. In Section 4.5.2 we define $\rho_S = 0.05$ and $\rho_C = 0$, implying the sensitivities are conditionally dependent but not the specificities; in Section 4.5.3 we reverse the relationship and define $\rho_S = 0$ and $\rho_C = 0.025$, implying indpendent sensitivities and dependent specificities. Finally, in Section 4.5.4 we generate X_1 and X_2 assuming dependent sensitivities and specificities $\rho_S = 0.05$ and $\rho_C = 0.025$.

For each generated dataset, we perform five analyses:

(1) we obtain posterior parameter estimates assuming both tests are independent

(2) we obtain posterior parameter estimates assuming the sensitivities are dependent and the specificities are independent

(3) we obtain posterior parameter estimates assuming the sensitivities are independent and the specificities are dependent

(4) we obtain posterior parameter estimates assuming both tests are dependent

(5) we estimate the DIC and the posterior model probability using Gibbs Variable Selection.

For each analysis, we place the following prior information:

$$\begin{aligned}
\beta &\sim N(\mathbf{0}, 20\mathbf{I}_4),\\
\lambda &\sim N(\mathbf{0}, 20\mathbf{I}_4),\\
\phi &\sim N(0, 20),\\
S_1 &\sim beta(1,1),\\
S_2 &\sim beta(1,1),\\
C_1 &\sim beta(1,1),\\
C_2 &\sim beta(1,1),
\end{aligned}$$

where $\mathbf{0}$ is a vector of length four with all elements equal to zero and $\mathbf{I}_4$ is the 4×4 identity matrix. Furthermore, for Sections 4.5.2 and 4.5.4 we define

$$\rho_S \sim uniform(L_S, U_S)$$

where $L_S = (S_1 - 1)(1 - S_2)$ and $U_S = min(S_1, S_2) - S_1S_2$. This induces a prior mean for ρ_S of 0.02 and a prior standard deviation of 0.029. Likewise for Sections 4.5.3 and 4.5.4 we define

$$\rho_C \sim uniform(L_C, U_C)$$

where $L_C = (C_1 - 1)(1 - C_2)$ and $U_C = min(C_1, C_2) - C_1C_2$. This induces a prior mean for ρ_C of 0.0125 and a prior standard deviation of 0.014. We note this as we analyze these parameters' posterior distributions to determine how substantially they differ from their prior distributions. For Sections 4.5.1 and 4.5.3 we assume $\rho_S = 0$, and for Sections 4.5.1 and 4.5.2 we assume $\rho_C = 0$.

We use R v2.7.2 to generate each dataset, and we subsequently use the R2WinBUGS program (Gelman et al., 2004) to estimate each model using WinBUGS v1.4.3 (Lunn et al., 2000). For each set of 800 observations, we performed a 1000 iteration burn-in followed by a 5000 iteration sample, thinning every third observation. Model convergence was assessed by visual inspection of the trace and autocorrelation plots.

4.5.1 Independent Binary Tests

For the first two tables, Tables 4.1 and 4.2, we generate data without dependence between tests and analyze the output using the four approaches explained in Section 4.5. The first column of posterior estimates matches the process by which the data were generated, and subsequently the posterior parameter means of the 200 datasets are reasonably close to the actual data generating value. This is unsurprising as Black and Craig (2002) effectively showed similar results, although with far more prior information required. However, it is of interest to determine the effect of attempting to estimate an effect that is not actually present within the data, especially given the constraints of a marginally identifiable model already.

The second column of Table 4.1 demonstrates the effect of estimating for dependent sensitivities when in fact they are independent. We note the addition of

ρ_S forces the observed sensitivities smaller, and subsequently the posterior mean for ρ_S is larger than that of the prior distribution, although its 95% credible interval contains zero and the standard deviation is nearly identical to the prior. We observe that ϕ, the model coefficient associated with the latent true classification status, is inflated. The mean posterior credible interval bounds still contain the "true" value 1.4, but compared to the model in which the dependence is ignored, there is clearly a loss of consistency in estimates. The disease model intercept (β_0) decreases while all remaining disease model posterior estimates ($\beta_1, \beta_2, \beta_3$) are attenuated. There is a universal increase in parameter standard deviations when we measure the model with dependent sensitivities, which reflects the added variability due to a superfluous parameter when identifiability is already strained.

In the first column of Table 4.2 we observe the posterior estimates when we attempt to estimate the model with an added dependent specificity parameter. Again, the posterior parameter means are not with the generating value, although in this case we observe attenuation in all disease model parameters except ϕ which is enlarged at approximately the same magnitude as when the sensitivities were modeled as dependent. All parameters in vector $\boldsymbol{\gamma}$ are biased away from zero, and the posterior means of the specificities are small.

In the final column of Table 4.2 demonstrates perhaps the most egregious of model misspecification, in which we assume both the sensitivities and specificities are dependent. The posterior means for all misclassification parameters (S_1, S_2, C_1, C_2) are biased, and this introduces a great deal of variability into the posterior distributions of the model parameters. The inconsistency demonstrated by modeling with a single superfluous dependence parameter is substantially exaggerated when both dependence parameters are included, and the posterior model parameter standard deviations are especially inflated. This model struggles to produce estimates similar to the data-generating distributions. In general, it is rather unsurprising that when the dual tests are independent, the analysis that assumes independence produces superior results and smaller parameter variances.

4.5.2 Conditionally Dependent Sensitivities

In Tables 4.3 and 4.4, we observe the posterior estimates for data generated with conditionally dependent sensitivities ($\rho_S = 0.05$) analyzed four ways. In the first analysis, we ignore the possible dependence and analyze the data with the two tests treated independently. When ρ_S is ignored, the mean posterior sensitivities are overestimated, but the specificities are unchanged. Despite the inconsistent estimates of the sensitivities, most of the posterior estimates of the model parameters are not dramatically affected. There is a slight attenuation of ϕ, but in general the other parameters in $\boldsymbol{\beta}$ are only mildly biased, although their 95% credible intervals contain the generating value.

In the second column on Table 4.3, we observe the posterior means for the data analyzed accounting for the dependent sensitivities. The 95% credible interval does not contain zero, and the posterior mean is substantially larger than the prior distribution mean, which allows us to conclude the model properly recognized the dependence parameter as being present in the data. However, the estimated sensitivities are mildly lower than the data generating values. Otherwise, the posterior model means are somewhat comparable to the independent model situation, although the standard devations are inflated. Given the similarities between the two means of estimation on Table 4.3, it does not appear that accounting for the dependence substantially improved the model parameter posterior estimates dramatically over the analysis assuming independence.

In Table 4.4, we analyze the data assuming dependent specificities. In this case, not only are the sensitivity posterior mean estimates too high, the specificity posterior mean estimates are too low due to the presence of a specificity dependence

Table 4.1. Data generated with conditionally independent tests, Part I

Analysis:		Assuming Independence			Dependent S_1 & S_2		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.693	0.175	(-1.044, -0.357)	-0.861	0.264	(-1.469, -0.426)
β_1	-0.3	-0.314	0.229	(-0.761, 0.134)	-0.216	0.282	(-0.720, 0.389)
β_2	0.8	0.805	0.233	(0.350, 1.263)	0.694	0.287	(0.090, 1.218)
β_3	1.6	1.604	0.281	(1.057, 2.159)	1.351	0.376	(0.560, 2.038)
ϕ	1.4	1.455	0.255	(0.967, 1.968)	1.732	0.397	(1.072, 2.636)
γ_0	-0.85	-0.894	0.237	(-1.373, -0.444)	-0.718	0.284	(-1.269, -0.154)
γ_1	-1.0	-1.064	0.398	(-1.905, -0.371)	-1.142	0.435	(-2.075, -0.399)
γ_2	1.0	1.025	0.271	(0.501, 1.565)	1.171	0.343	(0.556, 1.892)
γ_3	2.0	2.074	0.306	(1.496, 2.695)	3.037	1.257	(1.662, 6.414)
S_1	0.90	0.902	0.027	(0.847, 0.952)	0.834	0.056	(0.731, 0.938)
S_2	0.70	0.705	0.041	(0.626, 0.787)	0.634	0.060	(0.526, 0.755)
C_1	0.75	0.746	0.032	(0.684, 0.809)	0.743	0.032	(0.681, 0.806)
C_2	0.95	0.944	0.019	(0.905, 0.978)	0.941	0.019	(0.900, 0.976)
ρ_S	0	—	—	—	0.039	0.027	(-0.012, 0.087)
ρ_C	0	—	—	—	—	—	—

Table 4.2. Data generated with conditionally independent tests, Part II

Analysis:		Dependent C_1 & C_2			Dually Dependent S & C		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.681	0.181	(-1.044, -0.335)	-0.912	0.308	(-1.635, -0.426)
β_1	-0.3	-0.270	0.239	(-0.734, 0.202)	-0.093	0.341	(-0.677, 0.674)
β_2	0.8	0.755	0.249	(0.262, 1.237)	0.381	0.633	(-1.277, 1.179)
β_3	1.6	1.480	0.321	(0.829, 2.089)	0.665	1.013	(-2.064, 1.883)
ϕ	1.4	1.778	0.513	(1.062, 2.963)	2.636	1.162	(1.253, 5.754)
γ_0	-0.85	-1.123	0.342	(-1.876, -0.540)	-0.984	0.386	(-1.798, -0.285)
γ_1	-1.0	-1.835	1.350	(-5.511, -0.447)	-2.065	1.438	(-5.937, -0.500)
γ_2	1.0	1.144	0.332	(0.544, 1.853)	1.406	0.458	(0.631, 2.409)
γ_3	2.0	2.249	0.377	(1.582, 3.061)	3.538	1.359	(1.855, 7.041)
S_1	0.90	0.897	0.028	(0.839, 0.948)	0.810	0.056	(0.715, 0.924)
S_2	0.70	0.701	0.041	(0.622, 0.784)	0.611	0.060	(0.511, 0.739)
C_1	0.75	0.711	0.040	(0.635, 0.788)	0.699	0.039	(0.626, 0.778)
C_2	0.95	0.912	0.033	(0.845, 0.970)	0.899	0.033	(0.836, 0.962)
ρ_S	0	—	—	—	0.048	0.027	(-0.008, 0.091)
ρ_C	0	0.022	0.017	(-0.006, 0.057)	0.028	0.017	(-0.004, 0.061)

term. The models estimating $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and ϕ witness the most egregious deviations from the data generating values for γ_0, γ_1, and ϕ. However, these inconsistencies are comparable to those observed in the appropriate model (which assumes dependent sensitivities), with the exception of the poor sensitivity and specificity estimates. We conclude that to model conditionally dependent specificities when the sensitivities are actually dependent yields poor posterior estimates of the sensitivities and specificities while only slightly affecting the model parameter estimates.

The second column of Table 4.4 displays the estimates in which we attempt to account for dependent sensitivities and specificities in the presence of data generated with conditionally dependent sensitivities. Attempting to model dependence in both lacking substantial prior information yields diffuse posterior estimates that are not consistent with the data generating values. Furthermore, the sensitivities and specificities are far too small. We can conclude that attempting to model dually dependent misclassification probabilities lacking prior information yields poor model estimates.

4.5.3 Conditionally Dependent Specificities

We display the results for the third portion of the simulation in Tables 4.5 and 4.6 in which we generate 200 data sets with conditionally independent test sensitivities and conditionally dependent dual test protocol specificities, where $\rho_C = 0.025$. In the first column of Table 4.5 we present the posterior estimates of the analysis when we ignore the conditional dependence between the specificities and assume the tests are fully independent. It should be noted, due to the relatively high specificity of the second test where $S_2 = 0.95$, the magnitude of ρ_C is necessarily smaller than the sensitivity in the preceding Section 4.5.2. Due to this relatively small contribution of ρ_C, it comes as little surprise that the posterior estimates in which the dependence is ignored are fairly robust to the misspecification. The posterior

Table 4.3. Data generated with conditionally dependent sensitivities, Part I

Analysis:		Assuming Independence			Dependent S_1 & S_2		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.617	0.166	(-0.948, -0.296)	-0.820	0.276	(-1.462, -0.377)
β_1	-0.3	-0.353	0.223	(-0.789, 0.082)	-0.235	0.283	(-0.738, 0.376)
β_2	0.8	0.827	0.224	(0.390, 1.267)	0.717	0.273	(0.147, 1.222)
β_3	1.6	1.727	0.270	(1.205, 2.262)	1.466	0.362	(0.714, 2.133)
ϕ	1.4	1.296	0.232	(0.850, 1.757)	1.601	0.387	(0.966, 2.481)
γ_0	-0.85	-0.992	0.218	(-1.428, -0.574)	-0.732	0.304	(-1.298, -0.113)
γ_1	-1.0	-1.064	0.360	(-1.807, -0.406)	-1.156	0.403	(-2.004, -0.438)
γ_2	1.0	0.908	0.255	(0.413, 1.414)	1.071	0.342	(0.465, 1.796)
γ_3	2.0	1.759	0.267	(1.245, 2.291)	2.730	1.230	(1.424, 6.006)
S_1	0.90	0.966	0.018	(0.925, 0.995)	0.860	0.072	(0.728, 0.981)
S_2	0.70	0.777	0.044	(0.694, 0.865)	0.664	0.079	(0.525, 0.817)
C_1	0.75	0.745	0.030	(0.686, 0.805)	0.743	0.031	(0.682, 0.803)
C_2	0.95	0.949	0.015	(0.918, 0.977)	0.944	0.016	(0.910, 0.974)
ρ_S	0.05	—	—	—	0.062	0.036	(0.001, 0.121)
ρ_C	0	—	—	—	—	—	—

Table 4.4. Data generated with conditionally dependent sensitivities, Part II

Analysis:		Dependent C_1 & C_2			Dually Dependent S & C		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.602	0.170	(-0.941, -0.273)	-0.907	0.372	(-1.850, -0.379)
β_1	-0.3	-0.317	0.230	(-0.765, 0.134)	-0.078	0.391	(-0.685, 0.871)
β_2	0.8	0.788	0.234	(0.329, 1.244)	0.399	0.631	(-1.271, 1.181)
β_3	1.6	1.646	0.288	(1.082, 2.210)	0.812	0.972	(-1.793, 1.993)
ϕ	1.4	1.590	0.496	(0.929, 2.744)	2.515	1.176	(1.143, 5.680)
γ_0	-0.85	-1.248	0.349	(-2.044, -0.675)	-0.990	0.410	(-1.863, -0.237)
γ_1	-1.0	-1.892	1.404	(-5.744, -0.471)	-2.134	1.464	(-6.088, -0.541)
γ_2	1.0	1.037	0.333	(0.451, 1.763)	1.336	0.478	(0.549, 2.403)
γ_3	2.0	1.934	0.351	(1.322, 2.706)	3.295	1.387	(1.616, 6.870)
S_1	0.90	0.964	0.019	(0.921, 0.994)	0.829	0.074	(0.708, 0.969)
S_2	0.70	0.775	0.044	(0.691, 0.864)	0.633	0.079	(0.506, 0.798)
C_1	0.75	0.709	0.040	(0.632, 0.785)	0.698	0.039	(0.625, 0.775)
C_2	0.95	0.914	0.033	(0.846, 0.969)	0.901	0.032	(0.837, 0.960)
ρ_S	0.05	—	—	—	0.074	0.035	(0.005, 0.126)
ρ_C	0	0.025	0.019	(-0.006, 0.064)	0.031	0.019	(-0.003, 0.067)

mean specificities are slightly high, but few other posterior parameter means deviate substantially from the data generating values.

In the second column of 4.5, we observe the posterior estimates when we account for dependent sensitivities and ignore the dependent specificities. The misclassification probabilities S_1, S_2, C_1, and C_2 are inconsistent, and there is some noticeable bias in the estimated means within $\boldsymbol{\gamma}$. However, the disease model estimates $\boldsymbol{\beta}$ and ϕ fare generally well against this model misspecification.

In the first column of 4.6, we note that the model in which we correctly account for dependent specificities correctly excludes zero for the 95% credible interval for ρ_C. However, this produces slightly lower posterior mean specificities than the data generating value. Furthermore, there appears to be some instability in the estimates for $\boldsymbol{\gamma}$ and ϕ. Although this model matches the data generating scenario, it appears that the posterior distributions for the parameters are no better and perhaps worse than when we ignore the conditional dependence altogether. This may be due to the identifiability concerns we have already mentioned as well as the relatively small size of ρ_C.

The final column of 4.6 presents the posterior estimates when we attempt to account for conditional dependence in both sets of misclassification parameters. As in previous sections, the posterior mean misclassification probabilities are fairly distant from the data generating values, giving rise to diffuse and inconsistent posterior estimates for the model parameters. It again appears that the addition of a second model parameter for an already strained model results in poor posterior estimates.

4.5.4 Conditionally Dependent Sensitivities and Specificities

For the fourth simulation, we generate data with dually dependent sensitivities and specificities, $\rho_S = 0.05$ and $\rho_C = 0.025$. Table 4.7 displays the model posterior estimates when we ignore the dependence and assume conditionally independent

Table 4.5. Data generated with conditionally dependent specificities, Part I

Analysis:		Assuming Independence			Dependent S_1 & S_2		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.725	0.174	(-1.072, -0.391)	-0.845	0.228	(-1.336, -0.443)
β_1	-0.3	-0.308	0.225	(-0.748, 0.134)	-0.258	0.244	(-0.724, 0.232)
β_2	0.8	0.864	0.227	(0.420, 1.310)	0.808	0.247	(0.313, 1.282)
β_3	1.6	1.719	0.275	(1.189, 2.264)	1.585	0.314	(0.957, 2.187)
ϕ	1.4	1.280	0.223	(0.848, 1.722)	1.439	0.295	(0.911, 2.068)
γ_0	-0.85	-0.710	0.213	(-1.134, -0.301)	-0.563	0.261	(-1.056, -0.036)
γ_1	-1.0	-0.845	0.301	(-1.447, -0.272)	-0.883	0.316	(-1.519, -0.283)
γ_2	1.0	0.948	0.254	(0.456, 1.450)	1.055	0.314	(0.493, 1.707)
γ_3	2.0	1.948	0.288	(1.401, 2.529)	2.696	1.117	(1.519, 5.675)
S_1	0.90	0.905	0.022	(0.859, 0.947)	0.849	0.054	(0.743, 0.940)
S_2	0.70	0.700	0.039	(0.625, 0.779)	0.644	0.059	(0.533, 0.756)
C_1	0.75	0.789	0.033	(0.725, 0.853)	0.788	0.033	(0.724, 0.852)
C_2	0.95	0.978	0.013	(0.948, 0.997)	0.977	0.014	(0.946, 0.997)
ρ_S	0	—	—	—	0.033	0.028	(-0.014, 0.086)
ρ_C	0.025	—	—	—	—	—	—

Table 4.6. Data generated with conditionally dependent specificities, Part II

Analysis:		Dependent C_1 & C_2			Dually Dependent S & C		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.705	0.183	(-1.073, -0.353)	-0.936	0.322	(-1.697, -0.439)
β_1	-0.3	-0.228	0.241	(-0.696, 0.250)	-0.045	0.358	(-0.640, 0.760)
β_2	0.8	0.764	0.254	(0.257, 1.254)	0.336	0.694	(-1.490, 1.200)
β_3	1.6	1.499	0.339	(0.806, 2.129)	0.593	1.132	(-2.467, 1.934)
ϕ	1.4	1.877	0.670	(1.039, 3.522)	2.858	1.364	(1.243, 6.485)
γ_0	-0.85	-1.169	0.411	(-2.101, -0.498)	-1.054	0.432	(-1.995, -0.292)
γ_1	-1.0	-2.038	1.565	(-6.223, -0.365)	-2.258	1.636	(-6.612, -0.431)
γ_2	1.0	1.194	0.382	(0.544, 2.041)	1.468	0.498	(0.642, 2.580)
γ_3	2.0	2.291	0.430	(1.564, 3.251)	3.482	1.272	(1.832, 6.673)
S_1	0.90	0.895	0.025	(0.842, 0.941)	0.812	0.055	(0.717, 0.921)
S_2	0.70	0.694	0.040	(0.618, 0.774)	0.611	0.059	(0.511, 0.734)
C_1	0.75	0.717	0.046	(0.632, 0.809)	0.703	0.044	(0.624, 0.794)
C_2	0.95	0.912	0.039	(0.838, 0.980)	0.898	0.037	(0.829, 0.970)
ρ_S	0	—	—	—	0.074	0.035	(0.005, 0.126)
ρ_C	0.025	0.043	0.021	(0.006, 0.083)	0.050	0.021	(0.011, 0.087)

tests. We note that the misclassification probabilities are fairly substantially overestimated, including 95% credible intervals for S_1 and C_2 that do not contain the data generating values. Otherwise, the parameter estimates are affected although not as dramatically as the misclassification probabilities. We observe attenuation in the posterior means for $\boldsymbol{\gamma}$ which actually may help adjust for the overestimated misclassification parameters. Furthermore, the disease model parameters $\boldsymbol{\beta}$ and ϕ are fairly robust, with ϕ deviating perhaps more than any other parameter. This is likely due to the difficulty of estimation from biased S_1, S_2, C_1, and C_2.

In Table 4.7, we include a parameter to estimate the conditional dependence between the sensitivities while ignoring the dependent specificities. The posterior estimates from this model are generally more consistent with the data generating values than the independence model, which is likely due to the relatively large value of the sensitivity dependence parameter ρ_S. Once this parameter has been included, it appears most posterior estimates are subsequently adjusted and the exclusion of ρ_C is not deleterious to estimation.

We observe the effect of ignoring ρ_S and estimating ρ_C in the first columns of Table 4.8. While the specificities are representative of the data generating values, the sensitivities are high due to the exclusion of ρ_S. We observe fairly substantial deviations for the means of γ_0 and γ_1 from the parameter values, and furthermore the standard deviation for ϕ is considerably larger in this scenario than for the two previous estimates we observed. In general, the inclusion of specificity dependence did not produce estimates as consistent as when we included sensitivity dependence, although this is likely due to the relatively larger magnitude of ρ_S compared to ρ_C.

The final approach to analyzing the dually-dependent binary tests is in the second column of Table 4.8. The posterior 95% credible intervals for the dependence parameters ρ_S and ρ_C both exclude zero, which indicate the model was sensitive enough to detect their presence; however, the estimated means are larger than the data generating values. This produces sensitivity and specificity estimates that are too small relative to the generating values. Furthermore, the model posterior values are highly inconsistent, most notably ϕ which greatly exaggerates the effect of T on the response Y. Although the data are generated with conditionally dependent sensitivities and specificities, due to the identification difficulties with this model in the absence of prior information, either ignoring the dependence or estimating the dependence parameter of the largest magnitude may alleviate the bias of a conditionally dependent dual test protocol and produce posterior estimates that are potentially more representative of the parameter value than to attempt to model both misclassification dependence parameters.

Although in practice it may be plausible to assume that the sensitivities and/or the specificities of a dual test protocol may be conditionally dependent, direct estimation of the magnitude of the dependence is extremely difficult in the absence of prior information. In the preceding simulation, we have demonstrated the performance of the model parameter posterior distributions when we

(1) correctly identify independent dual binary tests,

(2) attempt to model dependence parameters in independent binary tests,

(3) correctly identify dependence between test sensitivities, specificities, or both,

(4) falsely assume dependent binary tests are independent.

Without overwhelmingly compelling reason to model test dependence, our simulations show that ignoring dependence or estimating a single dependence parameter is potentially superior to attempting to fit a marginally identifiable model with two dependence parameters. To some extent the specific parameter values we selected may account for this result, but in general, it appears that the model posterior estimates are relatively robust to incorrect under-parameterization than it is to correct

Table 4.7. Data generated with conditionally dependent sensitivities and specificities, Part I

Analysis:		Assuming Independence			Dependent S_1 & S_2		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.639	0.166	(-0.969,-0.319)	-0.807	0.245	(-1.352,-0.389)
β_1	-0.3	-0.374	0.221	(-0.808,0.057)	-0.304	0.247	(-0.772,0.198)
β_2	0.8	0.892	0.222	(0.461,1.329)	0.815	0.249	(0.312,1.289)
β_3	1.6	1.797	0.266	(1.284,2.326)	1.615	0.319	(0.973,2.223)
ϕ	1.4	1.168	0.210	(0.759,1.583)	1.391	0.313	(0.848,2.072)
γ_0	-0.85	-0.825	0.201	(-1.225,-0.438)	-0.588	0.285	(-1.111,-0.001)
γ_1	-1.0	-0.848	0.291	(-1.428,-0.289)	-0.903	0.313	(-1.535,-0.305)
γ_2	1.0	0.859	0.242	(0.388,1.337)	1.009	0.336	(0.436,1.725)
γ_3	2.0	1.650	0.254	(1.160,2.155)	2.534	1.194	(1.310,5.766)
S_1	0.90	0.970	0.014	(0.938,0.994)	0.873	0.070	(0.738,0.983)
S_2	0.70	0.773	0.042	(0.694,0.857)	0.674	0.076	(0.534,0.816)
C_1	0.75	0.784	0.031	(0.723,0.845)	0.783	0.031	(0.721,0.845)
C_2	0.95	0.980	0.010	(0.958,0.996)	0.978	0.011	(0.953,0.995)
ρ_S	0.05	—	—	—	0.059	0.037	(0.000,0.124)
ρ_C	0.025	—	—	—	—	—	—

Table 4.8. Data generated with conditionally dependent sensitivities and specificities, Part II

Analysis:		Dependent C_1 & C_2			Dually Dependent S & C		
Parameter	Value	Mean	St Dev	95% CI	Mean	St Dev	95% CI
β_0	-0.7	-0.625	0.172	(-0.968,-0.294)	-0.940	0.375	(-1.874,-0.402)
β_1	-0.3	-0.315	0.232	(-0.768,0.141)	-0.075	0.396	(-0.698,0.870)
β_2	0.8	0.838	0.236	(0.374,1.299)	0.410	0.697	(-1.493,1.248)
β_3	1.6	1.684	0.290	(1.114,2.251)	0.779	1.064	(-2.078,2.036)
ϕ	1.4	1.641	0.622	(0.886,3.182)	2.659	1.319	(1.110,6.142)
γ_0	-0.85	-1.207	0.378	(-2.068,-0.590)	-0.982	0.438	(-1.929,-0.197)
γ_1	-1.0	-1.905	1.464	(-5.845,-0.390)	-2.200	1.566	(-6.353,-0.463)
γ_2	1.0	1.041	0.345	(0.442,1.801)	1.373	0.516	(0.549,2.535)
γ_3	2.0	1.897	0.365	(1.270,2.702)	3.298	1.413	(1.570,6.933)
S_1	0.90	0.967	0.016	(0.932,0.992)	0.830	0.074	(0.710,0.968)
S_2	0.70	0.772	0.042	(0.691,0.856)	0.632	0.078	(0.507,0.792)
C_1	0.75	0.723	0.045	(0.636,0.812)	0.707	0.044	(0.627,0.797)
C_2	0.95	0.921	0.038	(0.843,0.985)	0.904	0.038	(0.834,0.975)
ρ_S	0.05	—	—	—	0.078	0.037	(0.006,0.132)
ρ_C	0.025	0.041	0.023	(0.003,0.086)	0.050	0.023	(0.007,0.090)

over-parameterization. Although ignoring dependence in a potentially dependent dual test protocol may seem negligent on the part of the statistician, we note that the model posterior estimates' consistency deteriorate as additional parameters are introduced into the model.

These conclusions are valid only in situations in which prior information for the dual tests does not exist. The strength of the Bayesian approach can allow for incorporation of prior data as demonstrated by Black and Craig (2002). Furthermore, when there exists evidence that binary tests may be conditionally dependent, in the absence of prior information about the tests it may be desirable to assume that both the sensitivities and specificities are dependent. In the context of this example, it may be best to model a single dependence parameter, preferably the misclassification dependence parameter with the largest possible value as determined by $max(\rho_S) = min(S_1, S_2) - S_1S_2$ and $max(\rho_C) = min(C_1, C_2) - C_1C_2$.

4.5.5 Model Assessment and Selection

For each data set within the preceding simulations, we additionally recorded the DIC for each model as explained in Section 4.4.2, and we fit the data set with a Gibbs Variable Selection procedure as explained in Section 4.4.1. The goal was to identify a statistical procedure that may allow us to consistently identify the presence or absence of dependence between binary test protocols. Ideally, the DIC should be minimized for the correct (i.e. data generating) model, and the posterior probability via GVS should be maximized. We present the mean posterior DIC values and model probability values in Table 4.9.

We identify the data generating model across the top of the table, and we quantify each method's posterior evaluation of the four candidate models. If the selection methods presented in Table 4.9 performed as advertised, we should observe the model with the smallest DIC and highest posterior probability to have matching column and row headings. This clearly is not the case. Overall the DIC method tends to uniformly prefer the independence model, while the GVS method uniformly prefers the dually dependent model. Ultimately our conclusion is that, to some extent, both models fail to discriminate the data generating model with consistency. This is likely due to the identifiability concerns already addressed in subsequent sections. The DIC is a measure of how well a given set of data fit the selected model under the assumption of a "correct" number of parameters. In our situation, introducing semi-identifiable parameters such as ρ_S and/or ρ_C are simply considered extra noise within the model due to the difficulty in achieving unique estimates. On the other hand, Gibbs Variable Selection using binary indicator covariates minimizes a model's posterior prediction variability under zero-one loss Kadane and Lazar (2004). Although it was developed as a means for regression variable selection and not as a means to attempt to identify parameters in scarcely identifiable models, the posterior results show that the inclusion of dependence parameter will likely yield lower variance posterior predictions, even in situations where no dependence exists.

A closer look at Table 4.9 reveals that the DIC criteria produces fairly consistent results favored toward the independent model regardless of the means by which the data were generated. On the other hand, although the GVS criteria tends to favor the full dependence model with the greatest posterior probability, we note that the independent model is selected with the highest probability when the data generating model used independent sensitivities and specificities. Likewise, the other three models achieved the highest probability within its row for the appropriate data generating model. Therefore, the row marginal probabilities show evidence that the GVS tended to agree with the data generating model, with the intolerable caveat that column marginal probabilities always favor the model with both dependence parameters included.

Table 4.9. Mean Posterior Model Evaluation Estimates of Simulated Datasets

	Indept		Dep. S		Dep. C		Dep. S & C	
Model	DIC	GVS	DIC	GVS	DIC	GVS	DIC	GVS
Independent	2936	0.127	2580	0.068	2722	0.056	2283	0.031
Dependent S	5997	0.232	7682	0.282	7132	0.079	8858	0.129
Dependent C	5056	0.201	5930	0.113	5849	0.289	6774	0.140
Dep. S & C	5717	0.441	7035	0.536	6193	0.576	7556	0.701

4.6 Epidemiological Example

To assess the model robustness under the four situations described in Section 4.5, we use a dataset published previously by Ren and Stone (2007a) to display the effects on the parameter estimates of modeling conditional dependence terms ρ_S and ρ_C. The original aims of the study were to examine the relationship between initial oxygenation status, T, and inpatient hospitalization status Y, among individuals with community acquired pneumonia (CAP). However, the measurement of true oxygenation status is fallible, and researchers rely instead on oxygenation status collected prospectively in the emergency department (X_1) and oxygenation status collected retrospectively via a medical chart review (X_2). We define $X_1 = 1$ or $X_2 = 1$ when hypoxemia is observed, which we define as pulse oximetry less than 90% or PO_2 less than 60 mm Hg. Furthermore, each participant is assigned to one of four Pneumonia Severity Index (PSI) score categories (Fine et al., 1997) that summarize the patient's pneumonia risk profile. The PSI is contained in vector $\mathbf{z}$. The data arise from the Emergency Department Community Acquired Pneumonia (EDCAP) Trial published in 2004 (Yealy et al., 2004).

We define our prior distributions identically to that used in Section 4.5. We ran a 2000 iteration burn-in followed by a 10000 iteration sample from the posteriors, and we assessed the chain convergence by visual inspection of autocorrelation and trace plots. We ran 3 chains in parallel, thinning every third observation. We note that the model is fairly sensitive to starting values of the chains.

In Table 4.10 we observe the posterior estimates for the data set analyzed four ways. We note the similarities in the posterior estimates when we assume the binary tests are independent and when we add single the conditional dependence parameter for the specificities ρ_C. This is likely due to the large specificities of both tests, which constrain the domain of ρ_C to a relatively small range. Subsequently, there is only a slight change in the overall specificity between the case where all dependence is ignored and the case when ρ_C is included. The consequences of this change would be that ϕ, the parameter associated with the true oxygenation status, will be somewhat elevated from its already high value, although its credible interval is greatly reduced.

We further note a great deal of agreement in the posterior estimates when we model conditional dependence between the sensitivities and for the dually dependent sensitivities and specificities model. It is noteworthy that adding the dual dependence parameters did not result in highly unstable posterior estimates as occurred in the simulation. However, the specificities are greatly reduced when the tests are assumed conditionally dependent. As we observe, the sensitivities (and in the second case the specificities) decrease by approximately 0.2 when modeled with dependence as the dependence parameter appears to be quite large. At this point it may be reasonable to solicit the advice of an expert; while a small shift in the sensitivity value may seem innocuous, a very large shift may produce estimates contrary to the known or assumed properties of the test, especially when specificity values sink below 0.5.

We include the DIC and posterior probability for each model, although as we observed in Section 4.5 these criteria may not be as informative as we might hope.

Table 4.10. Data example from Ren and Stone, 2007

	Independent		Dependent S		Dependent C			Dep. S & C	
	Mean	95% CI	Mean	95% CI	Mean	95% CI		Mean	95% CI
β_0	-0.66	(-1.03, -0.30)	-0.83	(-1.46, -0.37)	-0.65	(-1.03, -0.29)		-0.82	(-1.43, -0.36)
β_1	1.30	(0.82, 1.79)	1.36	(0.81, 2.01)	1.31	(0.83, 1.82)		1.37	(0.81, 2.01)
β_2	2.14	(1.58, 2.75)	2.00	(1.17, 2.76)	2.14	(1.55, 2.74)		1.98	(1.17, 2.72)
β_3	4.11	(3.12, 5.29)	2.75	(-1.71, 4.96)	4.09	(3.11, 5.28)		2.60	(-2.36, 4.89)
ϕ	4.65	(2.01, 10.12)	5.15	(2.20, 11.07)	5.36	(2.08, 11.17)		5.79	(2.32, 11.44)
γ_0	-2.46	(-3.25, -1.80)	-1.97	(-2.95, -1.06)	-2.52	(-3.44,-1.79)		-2.01	(-3.08, -1.10)
γ_1	0.32	(-0.59, 1.25)	0.31	(-0.70, 1.28)	0.28	(-0.80, 1.31)		0.27	(-0.83, 1.30)
γ_2	1.43	(0.67, 2.26)	1.59	(0.69, 2.60)	1.48	(0.64, 2.43)		1.62	(0.69, 2.70)
γ_3	2.50	(1.81, 3.32)	3.97	(1.99, 8.38)	2.56	(1.80, 3.47)		4.00	(2.00, 8.35)
S_1	0.687	(0.579, 0.808)	0.482	(0.319, 0.761)	0.683	(0.573, 0.805)		0.486	(0.317, 0.764)
S_2	0.771	(0.690, 0.849)	0.559	(0.391, 0.833)	0.759	(0.676, 0.838)		0.560	(0.387, 0.825)
C_1	0.985	(0.964, 0.998)	0.984	(0.964, 0.998)	0.981	(0.953, 0.997)		0.981	(0.957, 0.997)
C_2	0.938	(0.897, 0.975)	0.940	(0.898, 0.977)	0.930	(0.885, 0.969)		0.933	(0.890, 0.971)
ρ_S	—	—	0.084	(-0.033, 0.138)	—	—	—	0.080	(-0.036, 0.136)
ρ_C	—	—	—	—	0.007	(-0.0002,0.025)		0.007	(0.000, 0.022)
DIC	2021		8022		2692			7936	
GVS	0.194		0.243		0.253			0.310	

4.7 Discussion

Conditional dependence between binary diagnostic tests creates difficulties in model estimation when prior information is not known or available. The Bayesian approach to parameter estimation is powerful in that it does not require model identifiability in the same sense as frequentist estimation. However, prior information is necessary for consistent estimation of posterior means, and we have identified that dual test protocols with both dependent sensitivities and specificities result in poor estimates of the posterior under "noninformative" prior distributions. We are more likely to achieve consistent posterior estimates by ignoring one or both dependence parameters rather than to attempt to model them in a nonidentifiable model.

Based on the poor performance of DIC and GVS in this paper, future research should focus on further Bayesian methods for identifying conditionally dependent binary tests with minimal prior assumptions.

CHAPTER FIVE

Concluding Remarks

5.1 Future Work in the Area of Bayesian Variable Selection

The approach we developed in Chapter 2 was developed and demonstrated for a misclassified binary covariate in a logistic regression model when gold standard data is available. Further research should appeal to more general forms of models, such as the broader class of generalized linear models and nonlinear or semiparametric models. Additionally, as we have demonstrated in subsequent chapters, misclassification can be adjusted multiple ways, including double sampling, repeated binary testing, and dual test protocols. Future approaches to Bayesian variable selection should either consider alternate misclassification correction schemes or be general enough to not require specification of the type of misclassification.

The model we propose has satisfactory performance for moderately large data sets based on the results from the simulation. The method is designed for large sample sizes, although its performance for smaller samples should be scrutinized. Furthermore, as mentioned in Chapter 2, the model should be tested for both continuous covariates as well as covariate interactions. Additionally, this method should be considered in future research as a data reduction tool for scenarios for small n and large p.

5.2 Future Work in the Area of Repeated Binary Diagnostic Testing

The methods proposed in Chapter 3 are both straightforward Bayesian extensions of the existing pass/fail literature as well as the introduction of novel approaches to estimating misclassification variability. These methods make the basic assumption that the true classification status T is fixed with respect to time. Furthermore we developed the method under fairly basic considerations, such as fixed covariate values. Modifying this governing assumption about the effect of time upon the study outcome T may radically change the form of the problem into a change point problem. Additionally, further work should consider possible time-dependent covariates in the model to assess whether these may affect the probability of disease detection.

The approach described also could easily be extended to a multivariate setting. The form of the data we analyzed from Fujisawa and Izumi (2000) was such that combining individual observations into vectors was impossible. However, as the antigens we observed likely occur dependently to one another, the proposed method could benefit from the full individual antigen profile rather than considering observations independently.

5.3 Future Work in the Area of Dual Test Protocols

Our contributions in Chapter 4 include assessing the quality of the dependence assumptions regularly placed on dual test protocols as well as their impacts on other covariates when the misclassified variable is a model predictor. Based on our results, the more conservative approach appears to be to ignore dependence or include a single dependence parameter rather than risk the instability introduced by nonidentifiability when we introduce excessive parameters given the data. Because informative priors make parameter estimation much stronger in this case, further work should be done on the process of incorporating prior information into these types of models. For example, if our external validation set is comprised of different types of individuals or observations than the main study, we require methods for matching the data from the external study with that of the main study individuals. Otherwise we could be introducing biased data into our results.

5.4 Comments

This work introduces several new approaches to parameter estimation and model selection in the presence of binary misclassified data. Almost all of our approaches require moderately large data sets, although in Chapter 3 we demonstrate that our method works well with a sample of only 38 unique observations. Due to the inherent high variability of binary data, this is an obstacle that may be overcome in future studies with adequate prior information on the parameters. However, without prior data, specification of prior distributions must be taken with great care to adequately include the full potential parameter space of the model while not overstating our knowledge or supposed knowledge of the parameter. This debate over "noninformative" priors is adequately introduced and debated in Kass and Wasserman (1996). In no portion of this dissertation did we assume informative knowledge of prior distributions, though this is debatable simply on the grounds of the form of the priors which we did employ.

APPENDICES

APPENDIX A

WinBUGS Code for Described Methods

The following code was written for use in WinBUGS v1.4.3 Lunn et al. (2000). Each section indicates the portion of the text for which it is intended to be used.

A.1 Bayesian Variable Selection Method from Chapter 2

```
covbvs<-function()
{
for (i in 1:N){
# y is perfectly observed response (1= injured, 0 = uninjured)
y[i] ~ dbern(pi[i])

logit(pi[i]) <- beta[1] + beta[2]*x1[i]*g[1] + beta[3]*x2[i]*g[2]
            + beta[4]*tx3[i]*g[3]

tx3[i] ~ dbern(p[i])
logit(p[i]) <- lambda[1] + lambda[2]*x1[i]*g[4] + lambda[3]*x2[i]*g[5]

x3[i] ~ dbern(q[i])

q[i] <- tx3[i]*se[i] + (1-tx3[i])*(1-sp[i])

logit(se[i]) <- gamma[1] + gamma[2]*x1[i]*g[6] + gamma[3]*x2[i]*g[7]

logit(sp[i]) <- tau[1] + tau[2]*x1[i]*g[8] + tau[3]*x2[i]*g[9]
}

for (i in 1:4){
beta[i] ~ dnorm(0.0,1.0E-3)
}

for (i in 1:3){
gamma[i] ~ dnorm(0.0,1.0E-3)
tau[i] ~ dnorm(0.0,1.0E-3)
lambda[i] ~ dnorm(0.0, 1.0E-3)
}

for (k in 1:terms){
  index[k]<-pow(2,k-1)
  g[k] ~ dbern(0.5) #Prior on models
  }

  mdl<-1+inprod(g[ ], index[ ])

  for (s in 1:models){
  pmdl[s]<-equals(mdl,s) #Create model indicators
  }
}
```

A.2 Repeated Binary Diagnostic Test Code for Chapter 3

1.2.1 General Beta Priors Model

```
model
{
for (i in 1:8)
{
p3[i,1] <- pow(QN[i],2)*tau[i] + pow(1-QP[i],2)*(1-tau[i])
p3[i,2] <- 2*(QN[i]*(1-QN[i])*tau[i] + (1-QP[i])*QP[i]*(1-tau[i]))
p3[i,3] <- pow(1-QN[i],2)*tau[i] + pow(QP[i],2)*(1-tau[i])
n3[i]<-sum(x3[i,])
x3[i,1:3] ~ dmulti(p3[i,1:3],n3[i])

p4[i,1] <- pow(QN[i],3)*tau[i] + pow(1-QP[i],3)*(1-tau[i])
p4[i,2] <- 3*pow(QN[i],2)*(1-QN[i])*tau[i] + 3*pow(1-QP[i],2)*QP[i]*(1-tau[i])
p4[i,3] <- 3*QN[i]*pow(1-QN[i],2)*tau[i] + 3*(1-QP[i])*pow(QP[i],2)*(1-tau[i])
p4[i,4] <- pow(1-QN[i],3)*tau[i] + pow(QP[i],3)*(1-tau[i])
n4[i]<-sum(x4[i,])
x4[i,1:4] ~ dmulti(p4[i,1:4],n4[i])

p5[i,1] <- pow(QN[i],4)*tau[i] + pow(1-QP[i],4)*(1-tau[i])
p5[i,2] <- 4*(pow(QN[i],3)*(1-QN[i])*tau[i] + pow(1-QP[i],3)*QP[i]*(1-tau[i]))
p5[i,3] <- 6*(pow(QN[i],2)*pow(1-QN[i],2)*tau[i] + pow(1-QP[i],2)*pow(QP[i],2)*(1-
p5[i,4] <- 4*(QN[i]*pow(1-QN[i],3)*tau[i] + (1-QP[i])*pow(QP[i],3)*(1-tau[i]))
p5[i,5] <- pow(1-QN[i],4)*tau[i] + pow(QP[i],4)*(1-tau[i])
n5[i]<-sum(x5[i,])
x5[i,1:5] ~ dmulti(p5[i,1:5],n5[i])

QN[i] ~ dbeta(1,1)
QP[i] ~ dbeta(1,1)
tau[i] ~ dbeta(1,1)
}
}
```

1.2.2 Dirichlet Misclassification Priors Model

```
model
{
for (i in 1:n)
{
y[i] ~ dbin(p1[i],k[i])
p1[i] <- (1-theta[1])*t[i] + theta[2]*(1-t[i])
t[i] ~ dbern(tau)
}

r1 ~ dgamma(1,1)
r2 ~ dgamma(1,1)
r3 ~ dgamma(1,1)
rt <- r1 + r2 + r3

theta[1] <- r1/rt
theta[2] <- r2/rt
theta[3] <- r3/rt

tau ~ dbeta(1,1)
}
```

1.2.3 Hierarchical Dirichlet Misclassification Priors Model

```
model
{
for (i in 1:n)
{
y[i] ~ dbin(p1[i],k[i])
p1[i] <- (1-theta1[i])*t[i] + theta2[i]*(1-t[i])
t[i] ~ dbern(tau)
theta1[i] <- r1[i]/rt[i]
theta2[i] <- r2[i]/rt[i]
r1[i] ~ dgamma(alpha[1],1)
r2[i] ~ dgamma(alpha[2],1)
r3[i] ~ dgamma(alpha[3],1)
rt[i] <- r1[i] + r2[i] + r3[i]
}

alpha[1] ~ dgamma(1,1)
alpha[2] ~ dgamma(1,1)
alpha[3] ~ dgamma(1,1)
tau ~ dbeta(1,1)
}
```

1.2.4 Model for Fujisawa Dataset

```
model
{
for (i in 1:8)
{
p3[i,1] <- pow(QN[i],2)*tau[i] + pow(1-QP[i],2)*(1-tau[i])
p3[i,2] <- 2*(QN[i]*(1-QN[i])*tau[i] + (1-QP[i])*QP[i]*(1-tau[i]))
p3[i,3] <- pow(1-QN[i],2)*tau[i] + pow(QP[i],2)*(1-tau[i])
n3[i]<-sum(x3[i,])
x3[i,1:3] ~ dmulti(p3[i,1:3],n3[i])

p4[i,1] <- pow(QN[i],3)*tau[i] + pow(1-QP[i],3)*(1-tau[i])
p4[i,2] <- 3*pow(QN[i],2)*(1-QN[i])*tau[i] + 3*pow(1-QP[i],2)*QP[i]*(1-tau[i])
p4[i,3] <- 3*QN[i]*pow(1-QN[i],2)*tau[i] + 3*(1-QP[i])*pow(QP[i],2)*(1-tau[i])
p4[i,4] <- pow(1-QN[i],3)*tau[i] + pow(QP[i],3)*(1-tau[i])
n4[i]<-sum(x4[i,])
x4[i,1:4] ~ dmulti(p4[i,1:4],n4[i])

p5[i,1] <- pow(QN[i],4)*tau[i] + pow(1-QP[i],4)*(1-tau[i])
p5[i,2] <- 4*(pow(QN[i],3)*(1-QN[i])*tau[i] + pow(1-QP[i],3)*QP[i]*(1-tau[i]))
p5[i,3] <- 6*(pow(QN[i],2)*pow(1-QN[i],2)*tau[i] + pow(1-QP[i],2)*pow(QP[i],2)*(1-
p5[i,4] <- 4*(QN[i]*pow(1-QN[i],3)*tau[i] + (1-QP[i])*pow(QP[i],3)*(1-tau[i]))
p5[i,5] <- pow(1-QN[i],4)*tau[i] + pow(QP[i],4)*(1-tau[i])
n5[i]<-sum(x5[i,])
x5[i,1:5] ~ dmulti(p5[i,1:5],n5[i])

logit(QN[i]) <- fn + ref1[i]
logit(QP[i]) <- fp + ref2[i]
logit(tau[i]) <- b[1] + b[2]*equals(a[i],2) + b[3]*equals(a[i],3) + b[4]*equals(a
ref1[i] ~ dnorm(0,pr1)
ref2[i] ~ dnorm(0,pr2)
}
for (i in 1:4) {b[i] ~ dnorm(0,0.01)}
fn ~ dnorm(0,0.01)
fp ~ dnorm(0,0.01)
pr1 <- pow(sig1, -2)
pr2 <- pow(sig2, -2)
sig1 ~ dgamma(1,1)
sig2 ~ dgamma(1,1)
}
```

A.3 Code for Dependent Diagnostic Tests in Chapter 4

1.3.1 Dual Dependence Model for EDCAP Data

```
model
{
for (k in 1:n)
{
y[k]~dbern(py[k])

logit(py[k] ) <- bet[1]+bet[2]*tx[k]+bet[3]*equals(z[k],2)+bet[4]*equals(z[k],3)+

pf[k,1]<-tx[k]*(se1*se2+covse)+(1-tx[k])*((1-sp1)*(1-sp2)+covsp)
pf[k,2]<-tx[k]*(se1*(1-se2)-covse)+(1-tx[k])*((1-sp1)*sp2-covsp)
pf[k,3]<-tx[k]*((1-se1)*se2-covse)+(1-tx[k])*(sp1*(1-sp2)-covsp)
pf[k,4]<-tx[k]*((1-se1)*(1-se2)+covse)+(1-tx[k])*(sp1*sp2+covsp)

x[k,1:4] ~ dmulti(pf[k,1:4],1)

tx[k]~dbern(px[k])

logit(px[k]) <- gam[1]+gam[2]*equals(z[k],2)+gam[3]*equals(z[k],3) + gam[4]*equals
}

for (i in 1:5){
bet[i] ~ dnorm(0,0.05)
}
for(j in 1:4){
gam[j] ~ dnorm(0,0.05)
}

se1 ~ dbeta(1,1)
se2 ~ dbeta(1,1)
sp1 ~ dbeta(1,1)
sp2 ~ dbeta(1,1)

us <- min(se1,se2) - se1*se2
ls <- (se1 - 1)*(1-se2)
covse ~ dunif(ls, us)
uc <- min(sp1,sp2) - sp1*sp2
lc <- (sp1 - 1)*(1-sp2)
covsp ~ dunif(lc, uc)
}
```

1.3.2 Variable Selection Algorithm for EDCAP Data

```
model
{
for (k in 1:n)
{
y[k]~dbern(py[k])

logit(py[k] ) <- bet[1]+bet[2]*tx[k]+bet[3]*equals(z[k],2)+bet[4]*equals(z[k],3)+

pf[k,1]<-tx[k]*(se1*se2+g[1]*covse)+(1-tx[k])*((1-sp1)*(1-sp2)+g[2]*covsp)
pf[k,2]<-tx[k]*(se1*(1-se2)-g[1]*covse)+(1-tx[k])*((1-sp1)*sp2-g[2]*covsp)
pf[k,3]<-tx[k]*((1-se1)*se2-g[1]*covse)+(1-tx[k])*(sp1*(1-sp2)-g[2]*covsp)
pf[k,4]<-tx[k]*((1-se1)*(1-se2)+g[1]*covse)+(1-tx[k])*(sp1*sp2+g[2]*covsp)

x[k,1:4] ~ dmulti(pf[k,1:4],1)

tx[k]~dbern(px[k])
logit(px[k]) <- gam[1]+gam[2]*equals(z[k],2)+gam[3]*equals(z[k],3) + gam[4]*equals
}

for (i in 1:5){
bet[i] ~ dnorm(0,0.05)
}
for(j in 1:4){
gam[j] ~ dnorm(0,0.05)
}

se1 ~ dbeta(1,1)
se2 ~ dbeta(1,1)
sp1 ~ dbeta(1,1)
sp2 ~ dbeta(1,1)

us <- min(se1,se2) - se1*se2
ls <- (se1 - 1)*(1-se2)
covse ~ dunif(ls, us)
uc <- min(sp1,sp2) - sp1*sp2
lc <- (sp1 - 1)*(1-sp2)
covsp ~ dunif(lc, uc)

for (k in 1:2){
  index[k]<-pow(2,k-1)
  g[k] ~ dbern(0.5) #Prior on models
  }
mdl<-1+inprod(g[ ], index[ ])
for (s in 1:4){
pmdl[s]<-equals(mdl,s) #Create model indicators
  }
}
```

BIBLIOGRAPHY

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

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium of Information Theory*, eds. Petrov, B. N. and E, C., Akademiai Kiado., pp. 267–281.

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

Boyles, R. (2001), "Gauge Capability for Pass-Fail Systems," *Technometrics*, 43, 223–229.

Carlin, B. and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B*, 57, 473–484.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, vol. 105 of *Monographs of Statistics and Applied Probability*, Boca Raton, Florida: Chapman & Hall, 2nd ed.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Duxbury, 2nd ed.

Cheng, D., Stamey, J. D., and Branscum, A. J. (2007), "A General Approach to Sample Size Determination for Prevalence Surveys that Use Dual Test Protocols," *Biometrical Journal*, 5, 1–13.

Cochran, W. and Chambers, S. (1965), "The planning of observational studies of human populations," *Journal of the Royal Statistical Society, A*, 128, 234–255.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing*, 12, 27–36.

Dendukuri, N. and Joseph, L. (2001), "Bayesian Approaches to Modeling the Conditional Dependence between Multiple Diagnostic Tests," *Biometrics*, 57, 158–167.

Espeland, M. A. and Hui, S. L. (1987), "A General Approach to Analyzing Epidemiologic Data that Contain Misclassification Errors," *Biometrics*, 43, 1001–1012.

Fine, M. J., Auble, T. E.and Yearly, D., Hanusa, B., Weissfeld, L. A., Singer, D. E., Coley, C. M., Marrie, T. J., and Kapoor, W. N. (1997), "A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia," *The New England Journal of Medicine*, 336, 243–250.

Fujisawa, H. and Izumi, S. (2000), “Inference about Misclassification Probabilities from Repeated Binary Responses,” *Biometrics*, 56, 706–711.

Gelfand, A. E. and Ghosh, S. (1998), “Model choice: a minimum posterior predictive loss approach,” *Biometrika*, 85, 1–11.

Gelfand, A. E. and Smith, A. F. (1990), “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, 85, 398–409.

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

George, E. I. and McCulloch, R. E. (1993), “Variable Selection for Gibbs Sampling,” *Journal of the American Statistical Association*, 88, 881–889.

Gerlach, R. and Stamey, J. (2007), “Bayesian model selection for logistic regression with misclassified outcomes,” *Statistical Modelling*, 7(3), 255–273.

Greenland, S. (1989), “Modeling and Variable Selection in Epidemiologic Analysis,” *American Journal of Public Health*, 79, 340–349.

Gustafson, P. (2004), *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Boca Raton, Florida: Chapman & Hall.

Hochberg, Y. (1977), “On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors,” *Journal of the American Statistical Association*, 72, 914–921.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995), “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.

Kadane, J. B. and Lazar, N. A. (2004), “Methods and Criteria for Model Selection,” *Journal of the American Statistical Association*, 99, 279–290.

Kass, R. E. and Wasserman, L. (1996), “The Selection of Prior Distributions by Formal Rules,” *Journal of the American Statistical Association*, 91, 1343–1370.

Kullback, S. and Leibler, R. (1951), “On information and sufficiency,” *The Annals of Mathematical Statistics*, 22, 79–86.

Kuo, L. and Mallick, B. (1998), “Variable selection for regression model,” *Sankhya*, 60, Series B, Pt. 1, 65–81.

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

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

Mallows, C. L. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675.

Ntzoufras, I. (2002), "Gibbs Variable Selection Using BUGS," *Journal of Statistical Software*, 7.

Paulino, C., Soares, P., and Neuhaus, J. (2003), "Binomial Regression with Misclassification," *Biometrics*, 59, 670–675.

Powers, S., Stamey, J., and Gerlach, R. (2009), "Bayesian Variable Selection for Poisson Regression with Underreported Responses," Submitted for publication.

Prescott, G. J. and Garthwaite, P. H. (2005), "Bayesian Analysis of Misclassified Binary Data From a Matched Case Control Study With a Validation Sub-Study," *Statistics in Medicine*, 24, 379–401.

Qu, Y., Tan, M., and Kutner, M. H. (1996), "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests," *Biometrics*, 52, 797–810.

Reese, C. S., Deininger, P., Hamada, M. S., and Krabill, R. (2008), "Exploring the Statistical Advantages of Nondestructive Evaluation Over Destructive Testing," *Journal of Quality Technology*, 40, 259–267.

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

— (2007b), "A Bayesian approach for analyzing a cluster-randomized trial with adjustment for risk misclassification," *Computational Statistics and Data Analysis*, 51, 5507–5518.

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

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

Rogan, W. J. and Gladen, B. (1978), "Estimating prevalence from the results of a screening test," *American Journal of Epidemiology*, 107, 71–76.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 583–639.

Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Binomial Data with Misclassification," *Journal of the American Statistical Association*, 65, 1350–1361.

Vacek, P. M. (1985), "The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests," *Biometrics*, 41, 959–968.

van Wieringen, W. and van den Heuvel, E. (2005), "A Comparison of Methods for the Evaluation of Binary Measurement Systems," *Quality Engineering*, 17, 495–507.

Wang, J., Thornton, J., Russell, M., Burastero, S., S., H., and Pierson, R. (1994), "Asians Have Lower Body Mass Index but Higher Percent Body Fat Than Do Whites: Comparisons of Anthropometric Measurements," *American Journal of Clinical Nutrition*, 60, 23–28.

Yealy, D., Auble, T. E., Stone, R. A., Lave, J. R., Meehan, T. P., Graff, L. G., Fine, J. M., Obrosky, D. S., Edick, S. M., Hough, L. J., Tuozzo, K., and Fine, M. J. (2004), "The Emergency Department Community-Acquired Pneumonia Trial: Methodogy of a quality improvement intervention," *Annals of Emergency Medicine*, 43, 770–781.