Biostatistical Methods for HIV Monitoring and Prevention

Brett S. Hanscom

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

James P. Hughes, Chair

Susanne May, Chair

Elizabeth Brown

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Biostatistical Methods for HIV Monitoring and Prevention

Brett S. Hanscom

Co-Chairs of the Supervisory Committee:
James P. Hughes
Department of Biostatistics

Susanne May
Department of Biostatistics

Part I

Pooled-testing methods can greatly reduce the number of tests needed to identify failures in a collection of samples. Existing methodology has focused primarily on binary tests, but there is a clear need for improved efficiency when using expensive quantitative tests, such as tests for HIV viral load in resource-limited settings. We propose a matrix-pooling method which uses the EM algorithm to identify individual samples most likely to be failures. Simulation studies show that the proposed method can improve testing efficiency by a modest amount, and dramatically reduce the total number of testing rounds needed to identify all failures. In settings where the turn-around time for testing services is significant, the EM method can substantial time savings. The EM method does not perform as well when the measurements of interest are highly skewed, as is often the case with viral load concentrations.

We therefore propose a second method that accommodates situations where target quantities do not follow a normal distribution. This approach uses Markov Chain Monte Carlo (MCMC) sampling to identify the failure status of individual samples, and is highly flexible in terms of assumptions regarding the distributions of both target quantities and measurement error. This method is further extended to include covariate information. Simula-

tion studies show that the proposed method can substantially reduce turn-around time as compared to existing group-testing methods, particularly when covariate data is available that are highly predictive of failure. The proposed method did not perform as well when applied to a series of real datasets taken from actual pooled specimens. This may be due to measurement error variances for viral load testing that are much higher than anticipated.

Part II

Discordant partner studies are commonly used to quantify per-act infectivity rates for HIV. Statistical models used to estimate these rates depend on self-reported sexual activity, which is notoriously unreliable. The degree to which misreported sexual activity can affect infectivity estimates has not previously been reported. By using a basic transmission model in the context of measurement error, we show that infectivity estimates can be severely biased, and that the size and direction of bias depends on the underlying infectivity rate and on the mean number of sex acts occurring in the observation window. We show that by modifying the size of this window, in certain circumstances it is possible to avoid measurement-error bias without having to use a more sophisticated statistical model. We also show that, when misreporting is ignored, covariate parameter estimates can be biased, particularly when subgroup differences are large or when sexual frequency is heterogeneous.

Recent discordant partner studies have begun collecting sexual history data from both partners. We propose a latent-variable transmission model which incorporates data from both partners, and accounts for the mis-reporting of sex acts as well as non-overlapping recall periods. Bayes-MCMC methods are used to generate parameter estimates. Simulation studies demonstrate that the proposed method can dramatically reduce measurement-error bias for per-act infectivity estimation. We then apply this method to the Partners PrEP study dataset and show that HIV infectivity may be 25% higher than would be estimated with a naive model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

For Rachel, Max, Sadie, and Levi.

Chapter 1

INTRODUCTION

1.1 The Global HIV Epidemic

With over two million new infections and 1.5 million deaths every year, HIV remains one of the deadliest communicable infections in the world (UNAIDS, 2014; WHO, 2014b). Effective and readily available treatments in resource-wealthy countries have rendered HIV infection highly manageable for individuals with access to care (UNAIDS, 2014). However, millions of HIV-infected people in resource-poor nations, particularly in Sub-Saharan Africa, do not have access to treatment or prevention information and continue to transmit HIV to sexual partners and birth-children (UNAIDS, 2014; WHO, 2014a). With effective vaccines still years from development (NIAID, 2012), improved management of existing infections and continued evaluation of prevention strategies are critical to controlling the HIV epidemic.

This dissertation explores two distinct topics: (1) Using pooled serum testing to reduce the time and cost associated with viral load monitoring of HIV infected individuals receiving antiretroviral treatment, and (2) estimating per-act infectivity based on HIV serodiscordant partner studies. In both cases novel statistical methods are proposed and shown to provide more efficient and more accurate results than prior methods. Each chapter of this dissertation is written as a stand-alone research article and contains its own introduction and background material. A brief summary of each chapter is also included here.

1.2 Matrix-pooled testing with non-binary biological tests

Pooled-testing methods have long been known to reduce the number of tests needed to identify cases of disease in biological samples (Dorfman 1943). One drawback associated with these methods is that they typically involve lengthy, iterative procedures requiring long turnaround times. Another limitation is that the majority of existing pooling methods apply only to binary tests, i.e. tests that indicate the presence or absence of a biological agent. In the case of HIV monitoring, for HIV-positive individuals treated with anti-retroviral therapy (ART) the presence of HIV-1 virus is already known, however it is critically important to know whether the serum concentration of HIV virus has surpassed the "treatment-failure" threshold. Once this threshold has been breached, an infected individual is more likely to transmit the virus, and may be host to treatment-resistant viral strains. Individuals experiencing treatment failure are also at risk for transitioning to AIDS. Unfortunately, the cost of HIV viral load tests are prohibitively expensive in the resource-limited settings where they are needed most.

Chapters 2 and 3 present three new pooling methods that apply to continuous measurements of concentration. These methods are designed to reduce the total number of tests needed to identify failures and the number of testing iterations, thereby minimizing both cost and time.

The first approach is a non-statistical extension to the "Simple Search" matrix algorithm developed by May et al. 2010. The second approach considers the case where the target quantity follows a normal distribution, and the distribution of each individual sample is assumed to be independent conditional on whether it is a "failure" or "normal". In this scenario, pooled samples follow a multivariate normal distribution, and an Expectation-Maximization (EM) algorithm is employed to identify failures. Whereas EM is typically used to estimate parameters in missing-data situations, here we assume that the distribu-

tional parameters are known (or well estimated from previous data) and instead use the EM algorithm to discover the missing data itself.

The third approach considers cases where target quantities do not follow a normal distribution. This approach uses Markov Chain Monte Carlo (MCMC) sampling to identify the failure status of individual samples, and is highly flexible in terms of assumptions regarding the distributions of both target quantities and measurement error. This method is further extended to include covariate information that is available for each individual included in the pooled samples. All methods are general and can be applied in any setting where the test of interest yields a continuous measure of concentration.

1.3 Estimating per-act infectivity in HIV-1 sero-discordant partner studies where sex acts are reported with error

Per-contact infectivity is a fundamental measure associated any infectious disease. If estimable, this quantity is useful for both disease modeling and clinical trials. For example, HIV transmission prevention strategies can be evaluated in a clinical trial based on whether they successfully reduce per-act infectivity, and knowledge of factors associated with HIV infectivity can help guide patient counseling. Sero-discordant partner studies are the gold-standard method for estimating HIV per-contact (per sex act) infectivity. Such studies enroll monogamous couples with one HIV positive partner and one HIV negative partner, and these couples are followed prospectively, queried about their sexual activity, and tested for HIV transmission. Per-act infectivity can then be estimated from a simple transmission probability model. Unfortunately, self-reported sexual activity is known to be error prone, and it is also well known from the statistical literature that measurement error can significantly bias parameter estimates.

In Chapter 4 the potential impacts of sex-act misreporting are explored using theoretical

calculations and simulation studies. By making assumptions about the frequency of sexual activity in the population, and further assumptions about the shape and spread of the measurement-error distribution, it can be shown that ignoring measurement error in this context can lead to substantially biased parameter estimates. These biases are not limited to estimates of per-act infectivity, but extend to covariate parameter estimates as well.

Chapter 5 presents a novel method for modeling discordant partner study data data in cases where both partners provide self-reported sex act information. This method exploits the fact that when both partners report sex acts, better information is available about the true number of acts than when only only one partner reports. It also takes into account the fact that partners often provide sex-act information for time periods that are not the same. By using latent-variables and MCMC methods to estimate parameters, these models are highly flexible and can accommodate a variety of measurement error distributions, missing data, and non-overlapping time intervals. Simulation studies are used to show that this method can substantially reduce parameter estimation bias. The method is applied to data from the Partners in Prevention PrEP study.

Chapter 2

EFFICIENTLY IDENTIFYING FAILURES USING QUANTITATIVE TESTS, MATRIX-POOLING AND THE EM-ALGORITHM

2.1 Abstract

Pooled-testing methods can greatly reduce the number of tests needed to identify failures in a collection of samples. Existing methodology has focused primarily on binary tests, but there is a clear need for improved efficiency when using expensive quantitative tests, such as tests for HIV viral load in resource-limited settings. We propose a matrix-pooling method which, based on pooled-test results, uses the EM algorithm to identify individual samples most likely to be failures. Two hundred datasets for each of a wide range of failure prevalence were simulated to test the method. When the measurement of interest was normally distributed, at a failure prevalence level of 15.6% the EM method yielded a 47.3% reduction in the number of tests needed to identify failures (as compared to testing each specimen individually). These results are somewhat better than the reduction gained by using the Simple Search method (44.9%) previously published by May et al. (2010). However, the EM procedure was able to identify failures in just 2.6 testing rounds, on average, as compared to an average of 19.2 testing rounds required by Simple Search. In settings where the turn-around time for testing services is significant, the reduction in testing rounds provided by the EM method is substantial. Unfortunately the EM method does not perform as well when the measurements of interest are highly skewed, as is often the case with viral load concentrations.

2.2 Introduction

Pooled-testing methods can greatly reduce the number of tests needed to identify cases of disease in biological samples, particularly when disease prevalence is low. One drawback

associated with these methods is that they typically involve lengthy, iterative procedures requiring long turn-around times. For example, when pooling is not being used all specimens can be sent to a laboratory at the same time and each specimen is tested individually. When pooling is being used, portions of individual specimens are blended together and the resulting 'pools' are sent to the lab for testing. Based on the results of the initial test results, additional pools may be constructed and again sent for testing. After one or more of these testing rounds, selected individual specimens are sent to the lab and tested to confirm disease. Depending on the turn-around time for the lab, and depending on how many additional testing iterations are required by the pooled-testing algorithm, the overall time needed to identify all disease cases can be large. Lengthy time delays can be clinically important in terms of maintaining contact with diseased patients in resource-poor settings.

Another limitation is that the majority of existing pooling methods only apply to binary tests, i.e. tests that indicate only the presence or absence of a biological agent. In some instances the question of interest is not whether a substance is present, but whether the amount of that substance is higher that a certain threshold. When monitoring HIV-1 viral load in patients treated with ART, for example, the presence of HIV-1 virus is already known, and the questions is whether the concentration of HIV virus has surpassed a critical threshold. In resource limited settings the cost associated with viral-load tests is very high. If pooled testing methods can sufficiently reduce the costs associated with viral-load monitoring, it may become possible to introduce viral-load monitoring in resource limited settings. Ideally, we would like to identify "treatment-failure" cases quickly, without the need to carry out a lengthy pooled-testing algorithm.

Here we develop a pooled-testing method which accounts for the amount of a disease agent present in biological samples, and rapidly identifies specific samples that have surpassed a critical threshold. Our method combines 2-dimensional matrix pooling with the

EM algorithm, and iteratively tests individual samples. Simulation studies show that this approach can reduce the number of tests needed to identify failures, and dramatically reduces turn-around time. The method is general and can be applied in any setting where the test of interest yields a continuous measure of concentration.

2.3 Background

Pooled testing, also known as 'group testing', is a method that has been successfully used to reduce the cost of identifying disease cases (or failures) in a set of individuals (or items). Pooled testing is an intuitive method for saving time and money, and has broad application. The basic idea involves taking small portions of each specimen, mixing them together into one or more pools, and then testing the mixed pools. Assuming the test is sensitive, if a pool tests negative, all individual specimens in that pool must be negative, and the cost of multiple individual tests has been saved. If a pool test is positive, further testing must be performed to identify positive cases in that pool. Provided that a majority of pools yield negative tests, which will generally be true if prevalence is low, substantial cost savings can be achieved.

Dorfman (1943) first quantified the conditions under which pooled testing is useful in the context of binary (positive/negative) tests. By formulating a simple probability model, Dorfman was able to quantify the expected benefit of pooled testing, as well as to compute optimal pool configurations. Testing each of N individual specimens separately can be thought of as the baseline (expensive) approach, and pooled-testing methods are evaluated by the expected percent reduction in tests T required to identify all cases, i.e. $1 - E[T/N]$. Savings can be substantial, for example Dorfman found that under a prevalence of 1% and (optimal) pool size of 11, we can expect an 80% reduction in the number of tests performed. In general, the potential efficiency gain is larger in populations with lower disease (or failure) prevalence.

Since Dorfman's initial paper, numerous authors have proposed a wide variety of improvements and extensions to the basic group-testing idea. For example Phadarfod and Sudbury (1994) proposed using a matrix-pooling approach whereby specimens are arranged into a two-dimensional matrix, and the groups formed by combining samples from each row and each column are tested. Each specimen residing at the intersection of a positive row and positive column is then tested individually. An important advantage to matrix-pooling designs is that by testing both row and column pools each specimen is effectively tested twice, and thus the probability of false-negative samples can be reduced. Phadarfod and Sudbury (1994) showed that by implementing a simple square-array testing scheme the probability of a false-positive sample can be reduced by more than ten-fold in many practical scenarios.

Nearly all published results regarding pooled testing are based on binary testing. Individual samples either contain or do not contain a certain substance, and likewise a pool of individual samples either contains or does not contain that substance. Binary testing is common in biomedical settings, and the conceptual simplicity of binary tests lends itself well to pooled testing methods. There are instances, however, when we are interested in how much of a compound is present. When testing for lead in lake water, for example, one may expect to find a small amount of lead in any given water sample, but would only be concerned if the amount in any individual sample exceeded a certain threshold. Similarly, in the context of viral-load monitoring, all individuals are expected to have low levels of viral RNA in their blood, but we are only concerned when viral load becomes too high.

Quantitative tests produce more detailed results than binary tests, and as a result pooled quantitative tests can provide more information than binary tests. Exploiting this idea, May et al. (2010) developed a sequential testing algorithm based on matrix pooled samples for identifying ART failure among HIV patients. The algorithm takes its strength from the

fact that if a pool of specimens test positive on a binary test, there is no way to tell how many members of the pool are positive. If an individual from that pool is tested and turns out positive, it is still necessary to test the remaining members of the pool (perhaps by re-pooling them) to determine whether other members are positive as well. On the other hand, if a quantitative test is performed on one member of a positive pool, and the amount of test material found in that sample is enough to explain the amount of material observed in the pool, then no further testing of individuals in that pool is necessary. The search algorithm developed by May et al. (2010) is more efficient than other pooling methods at prevalence levels between about 4% and 25%.

We now propose a statistical approach that uses a matrix-pooling strategy to predict which individual samples are most likely to have surpassed the failure threshold. By making assumptions about the distribution of the target substance, and further using the EM algorithm to estimate key parameters in this distribution, we extend the method reported by May et al. (2010). The objective is to improve efficiency, defined as the percent reduction in the number of tests needed to identify all cases, and also reduce the turn-around time, defined as the number of sequential testing iterations required to identify all cases. This method can be used for any application where quantitative testing is performed. Although we use the term "failure" to indicate a concentration value above a critical threshold, in other settings a high concentration may indicate "success". The methodology is the same regardless of how a threshold breach is labelled.

2.4 Methods

2.4.1 Overview

Our approach to testing matrix-pooled specimens is an iterative algorithm that alternates between (1) estimating failure prevalence using Expectation-Maximization (EM), and (2)

testing the individual specimens that are most likely to contain failure-level concentrations. Test results from individual (non-pooled) specimens are then fed back into the EM procedure, prevalence is re-estimated, and further individual specimens are identified for testing. Once no further test candidates are found, the procedure is complete.

2.4.2 The Model

Starting with n^2 specimens arranged into an $n \times n$ array, we form $2n$ row and column pools, and test each pool for the target substance. For i=1...n (rows) and j=1...n (columns) let $\boldsymbol{Y} = \{y_{ij}\}$ be the unobserved concentration values for each specimen. Let $\boldsymbol{Z} = \{z_{ij}\}$ represent the failure status of each specimen, where $z_{ij} = 1$ if the ijth specimen is a failure and $z_{ij} = 0$ otherwise. Assume the z_{ij}s are iid $Bernoulli(p)$ where p is the (unknown) failure prevalence in the population. Assume that the target quantities are normally distributed conditional on z_{ij},

$$y_{ij}|z_{ij} \sim N(\lambda z_{ij} + \theta(1 - z_{ij}), \sigma_f^2 z_{ij} + \sigma_n^2(1 - z_{ij})) \tag{2.1}$$

where λ is the mean concentration among failures, θ is the mean concentration among non-failures ($\lambda > \theta$), σ_f is the standard deviation among failures and σ_n is the standard deviation among normals in the population. Note that since the conditional distribution of y_{ij} is assumed to be normal for both failures and non-failures, the range of possible values is the same for failures and non-failures. It is therefore possible to be a failure with a very low value of y_{ij}, or a non-failure with a high value of y_{ij}. Failure status is defined as belonging to one population or the other, which is slightly different than a typical clinical definition where failure might be defined based on a specific threshold.

The z_{ij}s are iid, and therefore the y_{ij}s are also independent. Thus if we let $\boldsymbol{y}$ represent

the values y_{ij} of the matrix $\boldsymbol{Y}$ arranged into a single column vector $(y_{11}, y_{12}, ..., y_{nn})^T$, then

$$\boldsymbol{y}|\boldsymbol{Z} \sim MVN_{n^2}(\nu(\boldsymbol{Z}), \boldsymbol{\Gamma}(\boldsymbol{Z})) \tag{2.2}$$

where

$$\nu(\boldsymbol{Z}) = \begin{pmatrix} \lambda z_{11} + \theta(1 - z_{11}) \\ \lambda z_{12} + \theta(1 - z_{12}) \\ \vdots \\ \lambda z_{nn} + \theta(1 - z_{nn}) \end{pmatrix} \tag{2.3}$$

and

$$\boldsymbol{\Gamma}(\boldsymbol{Z}) = \begin{pmatrix} \sigma_f^2 z_{11} + \sigma_n^2(1 - z_{11}) & 0 & \dots & 0 \\ 0 & \sigma_f^2 z_{12} + \sigma_n^2(1 - z_{12}) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_f^2 z_{nn} + \sigma_n^2(1 - z_{nn}) \end{pmatrix} \tag{2.4}$$

We can now represent the row and column pool concentrations as a function of the individual specimen concentrations as follows:

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} = \boldsymbol{A}\boldsymbol{y} + \epsilon \tag{2.5}$$

where

$$\boldsymbol{A} = \frac{1}{n}\begin{pmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & 1 & 1 & \dots & 1 \\ 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & \dots & 1 \end{pmatrix} \tag{2.6}$$

and

$$\epsilon \sim MVN_{2n}(\boldsymbol{0}, \tau^2 \boldsymbol{I}) \tag{2.7}$$

where $\boldsymbol{A}$ has dimension $2n \times n^2$, $\boldsymbol{I}$ is the $2n \times 2n$ identity matrix, and ϵ (measurement error) is a $2n \times 1$ vector of iid normal random variables with variance τ^2. The conditional joint distribution of the row and column-pool measurements is then

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} | \boldsymbol{Z} \sim MVN_{2n}(\mu(\boldsymbol{Z}), \boldsymbol{\Sigma}(\boldsymbol{Z})) \tag{2.8}$$

where

$$\mu(\boldsymbol{Z}) = \boldsymbol{A}\nu(\boldsymbol{Z}) \tag{2.9}$$

$$= \begin{pmatrix} \frac{1}{n}(\lambda \sum_{j=1}^{n} z_{1j} + \theta \sum_{j=1}^{n}(1 - z_{1j})) \\ \frac{1}{n}(\lambda \sum_{j=1}^{n} z_{2j} + \theta \sum_{j=1}^{n}(1 - z_{2j})) \\ \vdots \\ \frac{1}{n}(\lambda \sum_{i=1}^{n} z_{in} + \theta \sum_{i=1}^{n}(1 - z_{in})) \end{pmatrix} \tag{2.10}$$

$$= \begin{pmatrix} \frac{\lambda-\theta}{n} \sum_{j=1}^{n} z_{1j} + \theta \\ \frac{\lambda-\theta}{n} \sum_{j=1}^{n} z_{2j} + \theta \\ \vdots \\ \frac{\lambda-\theta}{n} \sum_{i=1}^{n} z_{in} + \theta \end{pmatrix} \tag{2.11}$$

$$= \frac{\lambda - \theta}{n} \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Z}^T \end{pmatrix} \mathbf{1}_n + \theta \mathbf{1}_{2n} \tag{2.12}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{\Gamma}(\boldsymbol{Z})\boldsymbol{A}^T + \tau^2 \boldsymbol{I} = \frac{1}{n^2} \times$$

$$\begin{pmatrix}
\sum_{j=1}^{n}(\sigma_f^2)^{z_{1j}}(\sigma_n^2)^{1-z_{1j}} & \dots & 0 & (\sigma_f^2)^{z_{11}}(\sigma_n^2)^{1-z_{11}} & \dots & (\sigma_f^2)^{z_{1n}}(\sigma_n^2)^{1-z_{1n}} \\
0 & \dots & 0 & (\sigma_f^2)^{z_{21}}(\sigma_n^2)^{1-z_{21}} & \dots & (\sigma_f^2)^{z_{2n}}(\sigma_n^2)^{1-z_{2n}} \\
\vdots & & \vdots & \vdots & & \vdots \\
0 & \dots & \sum_{j=1}^{n}(\sigma_f^2)^{z_{nj}}(\sigma_n^2)^{1-z_{nj}} & (\sigma_f^2)^{z_{n1}}(\sigma_n^2)^{1-z_{n1}} & \dots & (\sigma_f^2)^{z_{nn}}(\sigma_n^2)^{1-z_{nn}} \\
(\sigma_f^2)^{z_{11}}(\sigma_n^2)^{1-z_{11}} & \dots & (\sigma_f^2)^{z_{n1}}(\sigma_n^2)^{1-z_{n1}} & \sum_{i=1}^{n}(\sigma_f^2)^{z_{i1}}(\sigma_n^2)^{1-z_{i1}} & \dots & 0 \\
(\sigma_f^2)^{z_{12}}(\sigma_n^2)^{1-z_{12}} & \dots & (\sigma_f^2)^{z_{n2}}(\sigma_n^2)^{1-z_{n2}} & 0 & \dots & 0 \\
\vdots & & \vdots & \vdots & & \vdots \\
(\sigma_f^2)^{z_{1n}}(\sigma_n^2)^{1-z_{1n}} & \dots & (\sigma_f^2)^{z_{nn}}(\sigma_n^2)^{1-z_{nn}} & 0 & \dots & \sum_{j=1}^{n}(\sigma_f^2)^{z_{in}}(\sigma_n^2)^{1-z_{in}}
\end{pmatrix}$$

$$+T \tag{2.13}$$

where $T = \tau^2 \boldsymbol{I}$. In situations where we have differential measurement error between normals and failures, we instead let T be

$$diag\Bigg(\sum_{j=1}^{n}(\tau_f^2)^{z_{1j}}(\tau_n^2)^{1-z_{1j}}, \dots, \sum_{j=1}^{n}(\tau_f^2)^{z_{nj}}(\tau_n^2)^{1-z_{nj}}, \sum_{i=1}^{n}(\tau_f^2)^{z_{i1}}(\tau_n^2)^{1-z_{i1}}, \dots, \sum_{i=1}^{n}(\tau_f^2)^{z_{in}}(\tau_n^2)^{1-z_{in}}\Bigg) \tag{2.14}$$

where τ_f^2 is the measurement error variance associated with failures and τ_n^2 is the measurement error variance associated with normals.

The complete data likelihood (the joint distribution of $\boldsymbol{r}$, $\boldsymbol{c}$, and $\boldsymbol{Z}$) can be written as

the product of the conditional and marginal distributions as follows:

$$
\begin{aligned}
f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{Z}|\boldsymbol{\Theta}) &= f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{Z}, \boldsymbol{\Theta}) f(\boldsymbol{Z}|\boldsymbol{\Theta}) && (2.15)\\
&= (2\pi)^{-n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp(-\frac{1}{2}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))^T \boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z})))\\
&\quad \times p^{\Sigma z_{ij}}(1-p)^{\Sigma(1-z_{ij})} && (2.16)
\end{aligned}
$$

where $\boldsymbol{\Theta} = (\lambda, \theta, \sigma_f, \sigma_n, \tau, p)$, and Σ represents the summation over both i and j.

2.4.3 Implementing the EM Algorithm

The EM algorithm begins by maximizing the expected complete-data log likelihood over the unknown parameters. In the context of matrix pooling the number of observed data points is small (on the order of $2n$ for n between 6 and 12) and hence insufficient to estimate all six distributional parameters in this model. We therefore assume that all parameters except p can be well estimated using pre-existing data and scientific knowledge of testing devices. This seems especially reasonable if we imagine that the lab is doing repeated testing from the same population and can learn about the parameters over time. The expected complete-data log likelihood, conditional on the observed data, is

$$
\begin{aligned}
E[l(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{r},\boldsymbol{c})]|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] &= E[log(f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&= E[-nlog(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))^T\boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z})) \\
&\quad +\Sigma z_{ij}log(p) + \Sigma(1-z_{ij})log(1-p)|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&= -nlog(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}E[(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))^T\boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&\quad +log(p)\Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&\quad +log(1-p)(n^2 - \Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]) \qquad (2.17)
\end{aligned}
$$

Only the final two terms

$$log(p)\Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] + log(1-p)(n^2 - \Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]) \qquad (2.18)$$

are associated with the unknown parameter of interest p, and so, the expected complete-data likelihood, conditional on $\boldsymbol{r}$, $\boldsymbol{c}$ and parameter estimates is a linear function of

$$E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]. \qquad (2.19)$$

However, the marginal distribution of $\boldsymbol{r}$ and $\boldsymbol{c}$, $f(\boldsymbol{r},\boldsymbol{c}|\boldsymbol{\Theta})$ is not known and can only be computed by summing over all 2^{n^2} values of $\boldsymbol{Z}$, and hence it would be difficult to compute

$E[\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{\Theta}]$ analytically. Instead we estimate the expectation by summing over only the configurations of $\boldsymbol{Z}$ which could potentially yield the observed row and column values. All other configurations would have very low probability, and hence would contribute very little to the expectation equation.

2.4.4 Approximate expectation of $\boldsymbol{Z}$ (E Step)

For each row and column we first estimate the number of failures based on measured concentrations r_i and c_j and on our preliminary estimates $\hat{\lambda}$ and $\hat{\theta}$ of λ and θ. Let f_{r_i} be the approximate number of failures that would be expected to yield the observed value of r_i. Choose f_{r_i} so that

$$1/n(\hat{\lambda}(f_{r_i} - 1) + \hat{\theta}(n - f_{r_i} + 1) \leq r_i \leq 1/n(\hat{\lambda} f_{r_i} + \hat{\theta}(n - f_{r_i}), \tag{2.20}$$

unless

$$r_i \leq 1/n(\hat{\lambda} + \hat{\theta}(n - 1)) \tag{2.21}$$

in which case $f_{r_i} = 1$ if $r_i > (t^*/n)$ and $f_{r_i} = 0$ otherwise, where t^* is the failure threshold. For each column j, estimate the number of failures f_{c_j} the same way. This method will tend to overestimate the numbers of failures in each row and column, and, ignoring measurement error, will prevent any failures from being missed. Let $\{\boldsymbol{Z}'\}$ be the set of all $n \times n$ matrix configurations of failures and non-failures (1's and 0's) such that the number of failures in each row and column is less than or equal to f_{r_i} and f_{c_j}. Then compute (2.19) as

$$E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \quad = \quad \sum_{\{\boldsymbol{Z}'\}} \boldsymbol{Z}' f(\boldsymbol{Z}'|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}) \tag{2.22}$$

where

$$f(\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}) \approx f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}) / \sum_{\{\boldsymbol{Z}'\}} f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}'|\boldsymbol{\Theta}) \tag{2.23}$$

2.4.5 The M-Step

At each iteration of the E-M algorithm we then maximize the expected complete-data log likelihood over p. Differentiaing (2.17) by p, and defining $\hat{z}_{ij} = E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]_{ij}$, we have

$$\frac{\partial}{\partial p} E[l(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{r},\boldsymbol{c}])|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] = \frac{1}{p}\Sigma\hat{z}_{ij} - \frac{1}{1-p}(n^2 - \Sigma\hat{z}_{ij}) \tag{2.24}$$

which, after setting equal to 0 and solving for p gives

$$\hat{p} = \frac{1}{n^2}\Sigma\hat{z}_{ij}. \tag{2.25}$$

EM estimation proceeds as usual by choosing starting values $\boldsymbol{\Theta}_{(0)} = (\hat{\lambda}, \hat{\theta}, \hat{\sigma}, \hat{\tau}, p_{(0)})$, and estimating $E[log(f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}_{(0)}]$ by using the method described in Section 2.4.4. Then maximize over the unknown portion of the parameter space to get $\hat{\boldsymbol{\Theta}}_{(1)}$, re-estimate the expected log likelihood by estimating $E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\hat{\boldsymbol{\Theta}}_{(1)}]$, and so on, iterating until the parameter estimates converge. "Convergence" is defined as the iteration where $\hat{p}_{(n)}$ changes by less than 0.0001 from one iteration to the next.

2.4.6 Test individual specimens, and repeat EM

Once the EM algorithm has converged, identify the failure configuration $\boldsymbol{Z}'$ with the highest conditional probability, and test each individual specimen indicated in $\boldsymbol{Z}'$ by a 1. Let $\boldsymbol{y}^*$ be the $k \times 1$ vector of measurements for all individually tested specimens, and let $\boldsymbol{z}^*$ be the corresponding elements of $\boldsymbol{Z}$. Now repeat the EM algorithm by finding

$$E[log(f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}] \tag{2.26}$$

which again reduces to finding

$$E[\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}_{(0)}] = \sum_{\{\boldsymbol{Z}'\}} \boldsymbol{Z}' f(\boldsymbol{Z}'|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}_{(0)}) \tag{2.27}$$

$$\tag{2.28}$$

where $\boldsymbol{\Theta}_{(0)}$ is set equal to the final $\boldsymbol{\Theta}_{(n)}$ from the prior round of EM. From (2.4.2) we can derive that

$$\begin{pmatrix} \boldsymbol{y}^* \\ \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} |\boldsymbol{Z} \sim MVN_{2n+k}(\mu(\boldsymbol{Z})^*, \boldsymbol{\Sigma}(\boldsymbol{Z})^*) \tag{2.29}$$

where

$$\mu(\boldsymbol{Z})^* = \begin{pmatrix} \lambda \boldsymbol{z}^* + \theta(\boldsymbol{1}_k - \boldsymbol{z}^*) \\ \frac{\lambda - \theta}{n} \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Z}^T \end{pmatrix} \boldsymbol{1}_n + \theta \boldsymbol{1}_{2n} \end{pmatrix} \tag{2.30}$$

and $\boldsymbol{\Sigma}(\boldsymbol{Z})^*$ is straightforward expansion of (2.13). The set $\{\boldsymbol{Z}'\}$ now depends on the observed values $\boldsymbol{y}^*$. Define r_i' as the approximate concentration in the ith row pool if the target substance in the tested cells were removed:

$$r_i' = r_i - \frac{1}{n}\sum_{j=1}^{n} y_{ij}^*. \tag{2.31}$$

For the ith row, estimate the number of failures in untested cells by choosing f_{r_i}' such that

$$1/n(\hat{\lambda}(f_{r_i}' - 1) + \hat{\theta}(n - f_{r_1}' + 1) \leq r_i' \leq 1/n(\hat{\lambda} f_{r_i}' + \hat{\theta}(n - f_{r_1}'), \tag{2.32}$$

unless

$$r_i' \leq 1/n(\hat{\lambda} + \hat{\theta}(n-1)) \tag{2.33}$$

in which case $f_{r_i}' = 1$ if $r_i' > (t^*/n)$ and $f_{r_i}' = 0$ otherwise. Now the estimated number of failures in row i is

$$f_{r_i} = \sum_{j=1}^{n} 1_{[y_{ij}^* > t^*]} + f_{r_i}', \tag{2.34}$$

and the updated number of failures in each column is compute similarly.

Upon convergence of each subsequent round of the EM algorithm, identify the most probable configuration $\boldsymbol{Z}'$ and test any indicated specimens that have not previously been tested. If there are no further candidate specimens to test, the process is complete.

2.5 Simulation studies

In order to determine whether this methodology can reduce turn-around time and improve efficiency as compared to previously studied pooling methods, we simulate random samples of n^2 specimens for a variety of different values of n and prevalence p. We compare the EM method to the Simple Search method (May et al., 2010) and to a modified version of the Simple Search, comparing efficiency, turn-around time, Sensitivity, and Negative Predictive Value (NPV). "Efficiency" is defined as the percent reduction in the number of tests needed to detect all failures, as compared to testing each individual separately.

2.5.1 Comparable Methods

The "Simple Search" method developed by May et al. (2010) involves testing the row and column pools of a $n \times n$ matrix array and then testing the individual cell at the intersection of the row and column with the highest combined concentration that exceeds the failure threshold t divided by n. The tested value is then subtracted from the corresponding row and column pool values, and the cell with the next highest combined row and column concentration is tested. This process continues, testing one cell at a time, until there are no cells having both a row and column that exceed t/n.

The "Modified Simple Search" is a variation on the Simple Search method designed to reduce turn-around time. Whereas the Simple Search method identifies and tests the single cell with the highest combined row and column value, the Modified Simple Search further identifies and tests additional cells in unique rows and columns. Once the first cell is selected, the cell with the next highest combined row and column value is chosen from the remaining cells in distinct rows and columns. This process is repeated until approximately $n/2$ cells are identified for testing, and all $n/2$ cells are tested at once. Because more individual cells are identified for testing at each round, fewer testing rounds are necessary.

2.5.2 Normally distributed data

For the first set of simulation studies, target values were generated as a mixture of normal random variables with distinct means and variances, and with a common, small, measurement error variance. Two hundred simulated datasets were generated for each of a variety of fixed prevalence levels ranging from 0.01 to 0.16. Target values were simulated according to model (2.4.2), with $\lambda = 3000$, $\theta = 200$, $\sigma_f = 200$, $\sigma_n = 50$, and $\tau = 5$. The modelling procedure used the following parameter estimates: $\hat{\lambda} = 3100$ $\hat{\sigma}_f = 210$, $\hat{\theta} = 220$, $\hat{\sigma}_n = 48$ and $\hat{\tau} = 5$, with failure threshold $t^* = 1000$. Efficiency, turn-around time and sensitivity results are shown in Tables 2.1 and in Appendix A.

For prevalences less than about 6% all three methods perform similarly well in terms of efficiency. For prevalences above 10% the EM outperforms both Simple Search methods, with the advantage of EM increasing for higher prevalences. At the highest prevalence level tested (15.6%, or 10 failures in an 8×8 matrix), the EM method achieved 47.4% efficiency, while the SS and MSS methods were only 44.9% and 41.6% efficient, respectively.

In terms of turn-around time the EM method outperforms Simple Search methods even at low prevalances, and the benefit increases dramatically with increasing prevalence. At 15.6% prevalence the EM cuts the turn-around time in half as compared to the MSS, and improves on the SS method by 86%.

2.5.3 Skewed Data - HIV-1 Viral Load Example

In resource-wealthy settings, HIV-infected individuals who are being treated with antiretroviral medications (ART) are routinely monitored for virologic failure, defined as a detectable proliferation of HIV virus in the blood despite treatment. Individuals experiencing virologic failure may decline into worse health and experience AIDS, and they also may

Table 2.1: Simulation results for 8x8 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
Rounds - EM	1.0	1.3	1.4	1.7	1.9	2.1	2.4	2.4	2.5	2.6
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.1	2.5	3.2	3.9	4.4	5.0	5.7
Rounds - Simple Search	1.0	2.4	3.7	5.7	7.7	10.0	12.6	14.3	16.8	19.2
Efficiency - EM	73.4	71.1	68.9	66.1	62.9	59.8	56.4	54.1	50.8	47.4
Efficiency - Mod. Simple Search	73.4	69.3	67.7	64.3	61.2	57.1	52.6	49.4	45.8	41.6
Efficiency - Simple Search	73.4	71.3	69.2	66.2	63.0	59.4	55.3	52.6	48.8	44.9
Sensitivity - EM	100	100	100	100	100	100	99.7	100	100	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Individual Testing	100	100	100	100	100	100	99.9	100	100	100

present an increased risk of transmitting treatment-resistant virus to sexual partners. In the context of low-income countries, regular viral-load testing (performed by reverse transcriptase polymerase chain reaction (RT PCR)) is expensive and time consuming. Limited HIV funds must also be allocated to identifying HIV cases and providing treatment, education, prevention interventions, and a host of other services. It is critical therefore in these settings to minimize the cost of viral monitoring.

In general we do not expect true viral-load values to follow a normal distribution. Empirical data suggest that viral-load values tend to be skewed and are often well described by the lognormal distribution. In addition, assay measurement error tends to be constant on the log scale, with assay standard deviations of approximately 0.12 on the log_{10} scale (Brambilla et al., 1999), suggesting substantial variation for large viral loads. To test robustness to distributional assumptions, we also ran a simulation study where viral load values are not assumed to follow a normal mixture.

Simulated values for individuals with supressed viral load values were generated as exponential random variables with a mean of 50. Failure values were generated as lognormal with mean 3.2 and standard deviation 0.25 (on the log_{10} scale), and shifted by +1000, giving an actual mean of about 2,585. Measurement error was generated on the log_{10} scale as normal with mean 0 and standard error 0.12. Because measurement error is applied on the log_{10} scale, we use the second measurement-error variance term T as specified in (2.14), which accommodates the much larger measurement error among failures on the standard scale. When running the EM modeling procedure, the following parameter estimates were assumed: $\hat{\lambda} = 2800$, $\hat{\sigma}_f = 750$, $\hat{\tau}_f = 900$, $\hat{\theta} = 50$, $\hat{\sigma}_n = 50$ and $\hat{\tau}_n = 5$, with failure threshold $t^* = 1000$.

Two hundred datasets were generated for each prevalence level, with prevalence levels

fixed so that each dataset contained the same, known, number of failures. Simulation results are displayed in Table 2.2. Although turn-around times and efficiency look favorable for the EM method, these promising values come at the expense of a dramatic loss in sensitivity. All three methods show reduced sensitivity as compared with the normal-data simulations, however sensitivity for the the EM method is particularly low, at only 74% at the 16% prevalence level.

2.6 Discussion

These results suggest that for biological samples with normally distributed assay values, the EM-Matrix Pooling approach is quite effective. Although the cost savings associated with this approach is not substantial (about 3% at high prevalence levels), the time savings associated with shorter turn-around times is large (about 85% compared to Simple Search and 50% compared to Modified Simple Search). In cases where observed assay data is likely skewed such as HIV-1 viral load concentrations which tend to be lognormal, and particularly where measurement error is quite large, this method breaks down and does not yield high enough sensitivity to warrant use. The problem seems to be that the testing algorithm finishes too quickly, and is unable to recognize that additional, untested failures are present in the matrix. Perhaps it should not be surprising that large measurement error associated with pooled data would make inference about individual samples very difficult.

We did run simulations where the assay data was assumed to be lognormal, but the measurement error was small, and the the EM method performed well in terms of sensitivity and turn-around time. However, efficiency was not improved as compared to the Simple Search methods. (Results not shown.) Still, in situations where each round of testing requires a substantial amount of time, this method could be quite useful.

One limitation to our method is the requirement that most of the distributional parameters are known or well estimated. Although it would be nice to estimate or update these

Table 2.2: Simulation results for 8x8 matrix pools. Skewed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents) (Note - these results are not final. Simulations for higher prevalence levels were not completed (stopped due to futility). n for top three prevalence levels are 197, 84, 4.)

	Prevalence (%)									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
Rounds - EM	0.9	1.2	1.5	2.0	2.4	3.0	3.2	3.7	4.0	4.0
Rounds - Mod. Simple Search	1.0	1.0	1.6	2.2	2.8	3.5	4.0	4.7	5.3	6.0
Rounds - Simple Search	1.1	2.4	4.0	5.9	7.8	10.2	12.3	14.5	17.0	19.2
Efficiency - EM	73.5	71.2	68.5	65.3	62.2	58.5	56.0	52.4	50.3	50.0
Efficiency - Mod. Simple Search	73.1	69.4	66.8	63.2	59.5	54.9	51.6	47.5	43.9	39.5
Efficiency - Simple Search	73.3	71.3	68.7	65.8	62.7	59.0	55.8	52.3	48.4	44.9
Sensitivity - EM	89.5	90.5	90.2	87.8	87.2	84.8	80.0	77.5	74.7	62.5
Sensitivity - Mod. Simple Search	98.5	99.5	96.5	96.2	94.4	93.7	91.7	90.8	89.3	95.0
Sensitivity - Simple Search	98.5	96.5	94.5	94.2	93.6	92.2	90.9	90.2	89.0	92.5
Sensitivity - Individual Testing	99.5	100	98.7	98.9	99.7	99.2	99.0	99.3	99.2	100

parameters based on pooled data, it is unlikely that the very sparse data associated with matrix-pooled measurements would provide good information. However, in a laboratory setting where large numbers of biological samples are processed each day, it would be quite reasonable to estimate distributional parameters for populations that are sampled regularly over time. In addition, it would not be difficult to obtain repeated tests on certain samples in order to estimate measurement-error distributions. Assay calibration may require this.

Further investigation with normally distributed assay data should include an assessment of model performance under varying degrees of overlap between the failure distribution and non-failure distribution. The simulated examples reported here use distributions that have little overlap. An important limiting factor will be the need for the failure mean λ divided by n to be larger than the non-failure mean θ. If this is not the case, then a row with a single failure and the rest undetectable would be indistinguishable from a row of all non-failures with average concentrations. In general, it will likely be the case that for a given λ and θ, only pooling matrices of size n^* or less would be feasible, where n^* is the largest n for which $\lambda/n > c\theta$ for some constant c. A reasonable value of c would need to be investigated. Sensitivity to larger measurement error should also be explored.

In conclusion, this method was developed in the hope that HIV-1 viral load monitoring could be made faster and more efficient. Although the method works for normal data with minimal measurement error, unfortunately it does not seem to work well for skewed data with substantial measurement error, such as HIV-1 viral loads.

2.7 Appendix A

Table 2.3: Simulation results for 5x5 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	4	8	12	16	20	24	28	32	36	40
Rounds - EM	1.0	1.2	1.4	1.6	1.7	1.8	1.9	1.9	2.0	2.1
Rounds - Mod. Simple Search	1.0	1.2	1.9	2.3	2.9	3.5	4.1	4.6	5.1	5.5
Rounds - Simple Search	1.0	2.3	3.8	5.4	7.1	8.8	10.2	11.9	13.5	15.0
Efficiency - EM	56.0	50.4	44.9	38.8	32.9	26.9	21.8	17.2	10.3	5.1
Efficiency - Mod. Simple Search	56.0	48.6	42.5	35.1	28.6	21.4	14.5	8.4	2.7	-1.6
Efficiency - Simple Search	56.0	50.7	44.9	38.3	31.7	24.7	19.1	12.6	5.9	-0.0
Sensitivity - EM	100	100	100	99.9	100	100	100	100	99.8	100
Sensitivity - Mod. Simple Search	100	100	100	99.9	100	100	100	100	99.9	100
Sensitivity - Simple Search	100	100	100	99.9	100	100	100	100	99.9	100
Sensitivity - Individual Testing	100	100	100	99.9	100	100	100	100	99.9	100

Table 2.4: Simulation results for 6x6 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	2.8	5.6	8.3	11.1	13.9	16.7	19.4	22.2	25	27.8
Rounds - EM	1.0	1.1	1.4	1.6	1.8	1.9	2.0	2.2	2.2	2.3
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.0	2.5	3.0	3.6	4.1	4.5	4.9
Rounds - Simple Search	1.0	2.2	3.9	5.5	7.6	9.5	11.1	13.2	14.9	16.4
Efficiency - EM	63.9	60.3	55.9	51.4	46.6	43.2	38.1	32.7	29.0	25.2
Efficiency - Mod. Simple Search	63.9	57.3	53.6	48.4	42.3	36.7	30.5	24.7	20.5	16.3
Efficiency - Simple Search	63.9	60.4	55.9	51.4	45.7	40.4	35.8	30.1	25.3	21.2
Sensitivity - EM	100	100	100	99.5	100	99.8	99.7	100	99.6	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	99.9	99.9	100	99.9	100
Sensitivity - Simple Search	100	100	100	100	100	99.9	99.9	100	99.9	100
Sensitivity - Individual Testing	100	100	100	100	100	99.9	99.9	100	99.9	100

Table 2.5: Simulation results for 7x7 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	2	4.1	6.1	8.2	10.2	12.2	14.3	16.3	18.4	20.4
Rounds - EM	1.0	1.2	1.4	1.6	1.9	2.0	2.2	2.4	2.3	2.5
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.0	2.5	3.1	3.7	4.3	4.8	5.3
Rounds - Simple Search	1.0	2.3	3.8	5.7	7.9	9.7	12.1	13.9	16.0	18.1
Efficiency - EM	69.4	66.6	63.5	60.2	56.0	52.7	49.1	44.8	42.1	38.2
Efficiency - Mod. Simple Search	69.4	64.4	62.0	57.8	53.1	48.6	43.5	38.7	35.0	30.9
Efficiency - Simple Search	69.4	66.7	63.7	59.8	55.3	51.7	46.7	43.0	38.8	34.5
Sensitivity - EM	100	100	99.8	100	99.4	99.6	100	100	99.9	100
Sensitivity - Mod. Simple Search	100	100	99.8	100	99.8	99.9	100	100	99.9	100
Sensitivity - Simple Search	100	100	99.8	100	99.8	99.9	100	100	99.9	100
Sensitivity - Individual Testing	100	100	99.8	100	99.8	99.9	100	100	99.9	100

Table 2.6: Simulation results for 9x9 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1.2	2.5	3.7	4.9	6.2	7.4	8.6	9.9	11.1	12.3
Rounds - EM	1.0	1.2	1.5	1.8	2.0	2.2	2.4	2.6	2.8	2.9
Rounds - Mod. Simple Search	1.0	1.0	1.4	2.0	2.3	2.9	3.4	3.9	4.4	5.0
Rounds - Simple Search	1.0	2.4	3.9	6.0	7.9	10.4	13.1	15.3	17.4	20.1
Efficiency - EM	76.5	74.7	72.7	70.4	68.1	65.5	62.8	59.8	56.7	54.5
Efficiency - Mod. Simple Search	76.5	73.3	71.0	68.4	65.9	62.3	59.1	55.8	53.5	49.3
Efficiency - Simple Search	76.5	74.9	73.0	70.4	68.0	64.9	61.6	58.9	56.2	52.9
Sensitivity - EM	100	100	100	99.5	100	100	99.6	100	100	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Individual Testing	100	100	100	100	100	100	99.9	100	100	100

Table 2.7: Simulation results for 10x10 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1	2	3	4	5	6	7	8	9	10
Rounds - EM	1.1	1.2	1.5	1.9	2.1	2.4	2.6	2.7	2.9	3.0
Rounds - Mod. Simple Search	1.0	1.0	1.3	1.8	2.2	2.6	3.1	3.5	4.0	4.5
Rounds - Simple Search	1.1	2.5	4.1	6.1	8.2	10.5	13.2	15.8	18.3	20.3
Efficiency - EM	78.9	77.6	75.9	74.0	72.0	69.8	66.9	65.1	62.6	60.3
Efficiency - Mod. Simple Search	78.3	75.6	73.5	71.8	69.2	66.7	63.8	61.5	57.9	55.3
Efficiency - Simple Search	78.9	77.5	75.9	73.9	71.8	69.5	66.8	64.2	61.7	59.7
Sensitivity - EM	100	90.0	90.0	94.6	95.2	97.8	99.6	100	100	100
Sensitivity - Mod. Simple Search	100	100	99.8	100	99.9	100	100	100	100	100
Sensitivity - Simple Search	100	100	99.8	100	99.9	100	100	100	100	100
Sensitivity - Individual Testing	100	100	99.8	100	99.9	100	100	100	100	100

Chapter 3

EFFICIENTLY IDENTIFYING FAILURES USING QUANTITATIVE TESTS, MATRIX-POOLING AND A STATISTICAL APPROACH

3.1 Introduction

Pooled-testing methods can greatly reduce the number of tests needed to identify cases of disease in a population of individuals, particularly when disease prevalence is low. One drawback associated with these methods is that they can involve lengthy, iterative procedures requiring long turn-around times. Another limitation is that the majority of existing pooling methods apply only to binary tests, i.e. tests that indicate the presence or absence of a biological compound. In some applications the question of interest is not whether a substance is present, but whether the amount of that substance is higher than a certain level. For example, when monitoring HIV-1 viral load in patients treated with anti-retroviral therapy (ART), the presence of HIV-1 virus is already known, and it is of interest to know whether the concentration of HIV virus has surpassed a critical"treatment-failure" threshold.

Here we present two new pooling methods that apply to continuous measurements of concentration. These methods aim to reduce the total number of tests needed to identify failures, and also reduce the number of testing iterations, thereby minimizing both cost and time. The first is an extension of what we call the "Simple Search" matrix algorithm developed by May et al. (2010), and the second is a statistical approach. Both methods are general and can be applied in any setting where the test of interest yields a continuous measure of concentration.

3.2 Background

Pooled testing, also known as 'group testing', is a method that has been successfully used to reduce the cost of identifying disease cases (or failures) in a set of individuals (or items). Pooled testing is intuitively appealing and has broad application. The basic idea involves taking a biological specimen from each individual in the target population. A portion of each specimen is then mixed together into one or more pools, and the mixed pools are tested for the target substance. Assuming the test is sensitive, if a pool tests negative, all individual specimens in that pool must be negative, and the cost of multiple individual tests has been saved. If a pool test is positive, further testing must be performed to identify positive cases in that pool. Provided that a majority of pools yield negative tests, which will generally be true if prevalence is low, substantial cost savings can be achieved.

Dorfman (1943) first quantified the conditions under which pooled testing is useful in the context of binary (positive/negative) tests. By formulating a simple probability model, Dorfman was able to quantify the expected benefit of pooled testing, as well as to compute optimal pool configurations. Testing each of N individual specimens separately can be thought of as the baseline, expensive approach. One aspect by which pooled-testing methods can be evaluated is by the expected percent reduction in tests T required to identify all cases, i.e. $1 - E[T/N]$. Savings can be substantial; for example, Dorfman found that under a prevalence of 1% and (optimal) pool size of 11 we can expect an 80% reduction in the number of tests performed. In general, the potential efficiency gain is larger in populations with lower disease (or failure) prevalence.

Since Dorfman's initial paper, numerous authors have proposed a wide variety of improvements and extensions to the basic group-testing idea (Sterrett, 1957; Thompson, 1962; Sobel and Elashoff, 1975; Tanner and Wong, 1987; Litvak, 1994; Phadarfod and Sudbury, 1994; Brookmeyer, 1999; Xie et al., 2001). Phadarfod and Sudbury (1994) proposed using

a matrix-pooling approach whereby specimens are arranged into a two-dimensional matrix, and the groups formed by combining samples from each row and each column are tested. Each specimen residing at the intersection of a positive row and positive column is then tested individually. An important advantage to matrix-pooling designs is that by testing both row and column pools each specimen is effectively tested twice, and thus the probability of false-negative samples can be reduced. Phadarfod and Sudbury (1994) showed that by implementing a simple square-array testing scheme the probability of a false-negative sample can be reduced by more than ten-fold in many practical scenarios.

Nearly all published results regarding pooled testing are based on binary testing. Individual samples either contain or do not contain a certain substance, and likewise a pool of individual samples either contains or does not contain that substance. Binary testing is common in biomedical settings, and the conceptual simplicity of binary tests lends itself well to pooled testing methods. There are instances, however, when we are interested in how much of a compound is present. When testing for lead in lake water, for example, one may expect to find a small amount of lead in any given water sample, but would only be concerned if the amount in any individual sample exceeded a certain threshold. Similarly, in the context of viral-load monitoring, all individuals are expected to have low levels of viral RNA in their blood, but we are only concerned when viral load becomes too high.

Quantitative tests produce more detailed results than binary tests, and as a result pooled quantitative tests can, in theory, provide more information than binary tests. Exploiting this idea, May et al. (2010) developed a sequential testing algorithm based on matrix pooled samples for identifying ART failure among HIV patients. This “Simple Search” algorithm takes its strength from the fact that if a pool of specimens test positive on a binary test, there is no way to tell how many members of the pool are positive. If an individual from that pool is tested and turns out positive, it is still necessary to test the remaining members

of the pool (perhaps by re-pooling them) to determine whether other members are positive as well. On the other hand, if a quantitative test is performed on one member of a positive pool, and the amount of test material found in that sample is enough to explain the amount of material observed in the pool, then no further testing of individuals in that pool is necessary. The search algorithm developed by May et al. (2010) is shown to be more efficient than dichotomous-test pooling methods at prevalence levels between about 4% and 25%.

Hanscom et al. (2014a) proposed a statistical approach to the analysis of matrix pooled data that uses the Expectation Maximization (EM) algorithm (Dempster et al., 1977), and that method was shown to be effective when the quantities of interest are normally distributed. In many important applications such as HIV viral-load testing, however, the target quantities tend to follow asymmetric distributions such as the Lognormal, in which case the EM approach was shown to have poor sensitivity.

We now propose two new methods that exploit the information available in quantitative tests, the Modified Simple Search (MSS) and the Model Based Search (MBS).

3.3 Methods

3.3.1 Modified Simple Search (MSS)

The "Simple Search" (SS) matrix method developed by May et al. (2010) involves testing the row and column pools of an $n \times n$ matrix array and then identifying all pools that contain concentrations larger than the failure threshold t divided by n. The individual cell with the highest combined row and column concentration is then tested. The resulting quantity is then divided by n and subtracted from the corresponding row and column pool values, and the cell with the next highest combined row and column concentration is tested. This process continues, testing one cell at a time, then adjusting the row and column pool values, until there are no cells having both a row and column value that exceed t/n. See

Figure 3.1 for an illustration. In general we do not expect our target population to have exactly n^2 individuals, but rather we imagine a laboratory environment where samples are continually arriving for testing, and each time a batch of n^2 samples is accumulated, matrix testing can proceed.

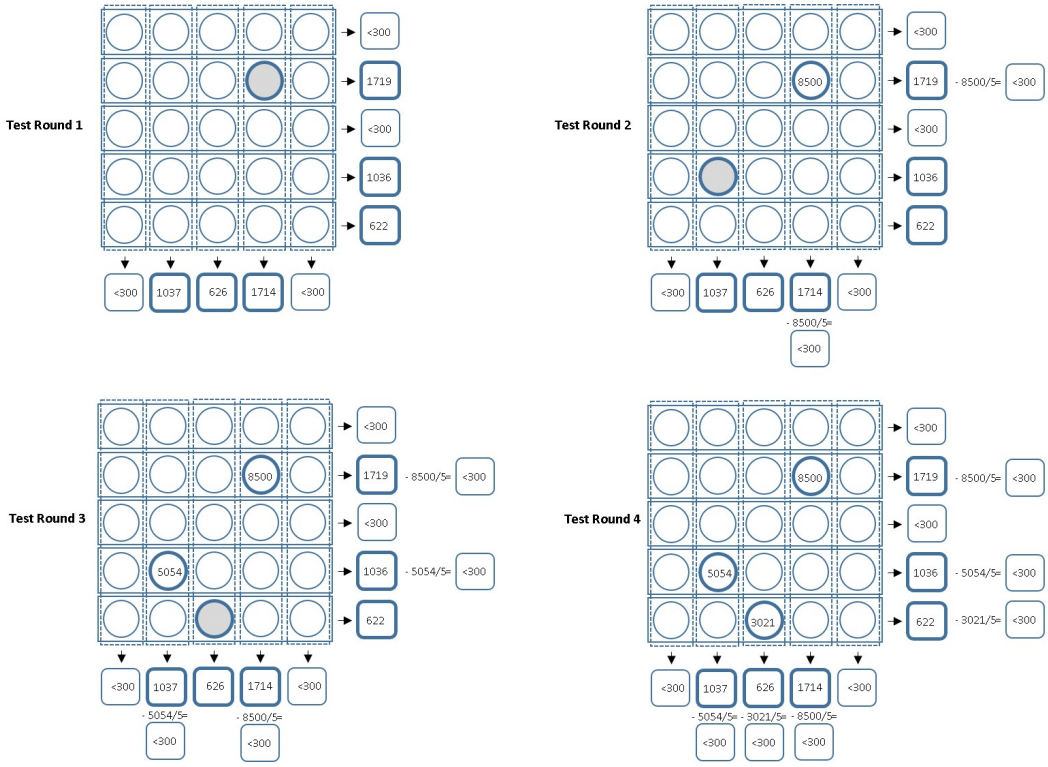

Figure 3.1: Schematic example of the Simple Search (SS) algorithm for a 5x5 matrix. Circles represent individual samples, and squares in the margins represent pooled samples. The failure thresthold is defined as 1500 RNA copies/ml. In round 1 the row and column pools are tested and the cell with the highest combined row and column concentration is identified (row 2, column 4). In round 2 that cell is tested, it's values is subtracted from the corresponding pool values, and the cell with the new highest combined row and column values is identified (row 4 column 2). In round 3 cell (4,2) is tested, its value subtracted out, and cell (5,3) is identified for testing. In round 4 cell (5,3) is tested, its value is subtracted out, and no further cells are identified for individual testing. A total of 13 RNA tests and 4 testing rounds were necessary to identify all failures.

The Modified Simple Search (MSS) is a variation on the Simple Search that is designed

to reduce the number of testing iterations and hence reduce turn-around time. MSS starts off in the same manner as SS by identifying all cells with corresponding row and columns pools that exceed t/n, and selecting the cell with the largest row and column pool sum. Once the first cell is selected, the cell with the next highest row and column pool sum is chosen from the remaining cells in distinct rows and columns. This process is repeated until $\lceil n/2 \rceil$ cells are identified for testing, or until there are no more cells in distinct rows and columns with row and column pools larger than t/n. See Figure 3.2 for an illustration using the same initial pooled values as in Figure 3.1. If there are fewer than $\lceil n/2 \rceil$ candidate cells in distinct rows and columns, then the requirement of unique rows and columns is relaxed and additional cells are chosen based on the sum of their row and column pools (largest to smallest). Again the selection process continues until a total of $\lceil n/2 \rceil$ cells have been selected for testing or until there are no additional cells with row and column pools exceeding t/n. Once the selection process is complete, all cells identified for testing are tested at the same time.

As in the SS, individually testing values are then divided by n and subtracted from their corresponding row and column pool values. The newly adjusted pool values are used for the next round of testing. Additional cells are identified for testing using the same procedure as in the first round, and the process of testing new batches of individual samples continues until there are no longer any cells with adjusted row and column pool values larger than t/n. Since multiple cells will typically be identified for testing at each round, fewer testing rounds are necessary and turn-around time can be reduced. Like the Simple Search, this method makes no assumptions regarding the distribution of the target concentration, and does not account for measurement error.

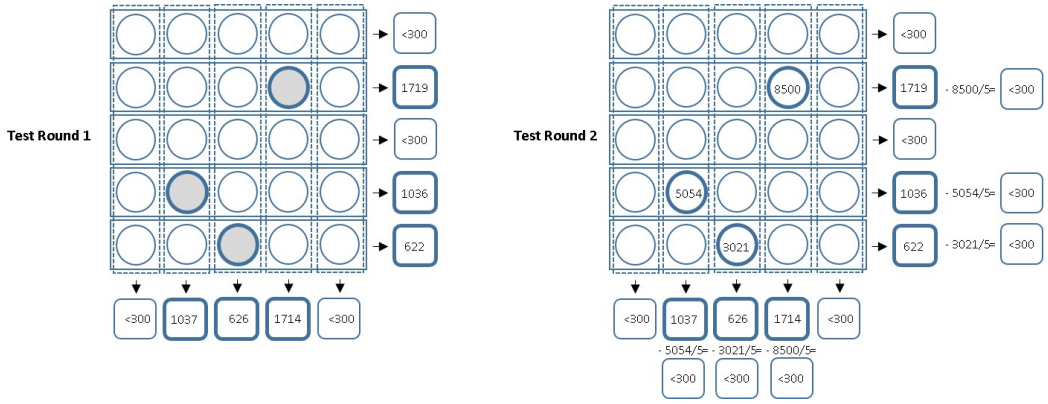

Figure 3.2: Schematic example of the Modified Simple Search (MSS) method used for the same samples as in Figure 3.1. In round 1 all row and column pools are tested and $\lceil n/2 \rceil = 3$ individual cells in unique rows and columns are all identified for individual testing. In round 2 all three cells are tested, their values are subtracted from the corresponding row and column pool values, and no further cells are identified for testing. A total of 13 RNA tests and 2 testing rounds were needed to identify all failures using the MBS.

3.3.2 Model Based Search (MBS)

The Model Based Search is an iterative algorithm that alternates between (1) using MCMC sampling to estimate the failure status of each individual, conditional on pooled-sample measurements, and (2) testing the individual specimens that are most likely to contain failure-level concentrations. Test results from individual (non-pooled) specimens are fed back into the MCMC sampling procedure, prevalence and failure status are re-estimated, and further individual specimens are identified for testing. Once no further test candidates are found, the procedure is complete. The MBS method assumes that row and column pool values follow a mixture distribution depending on the number of failures and non-failures in each pool, and also assumes that all pooled and individual test values are measured with error. Also, covariates can be incorporated into the model in order to help improve performance. Strong, pre-existing information regarding model parameters is leveraged by using informative prior distributions in a Bayesian framework.

Starting with n^2 specimens arranged into an $n \times n$ array, we form n row pools and n column pools, and test each pool for the target substance. For i=1...n (rows) and j=1...n (columns) let $\boldsymbol{Y} = \{y_{ij}\}$ be the unobserved concentration values for each specimen. Let $\boldsymbol{Z} = \{z_{ij}\}$ represent the failure status of each specimen, where $z_{ij} = 1$ if $y_{ij} > t$ for the ijth specimen and $z_{ij} = 0$ otherwise, and where t is the failure threshold. Both $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are unobservable latent or "auxiliary" variables used to construct an estimable model, as described by Tanner and Wong (1987).

We assume the z_{ij}'s are iid $Bernoulli(p)$ where p is the (unknown) failure prevalence in the population. If covariates $\boldsymbol{x}$ are available with known association with failure, then let $z_{ij} \sim Bernoulli(p_{ij})$ where

$$\text{logit}(p_{ij}) = \beta_0 + \boldsymbol{x}_{ij}\boldsymbol{\beta}. \tag{3.1}$$

where β_0 is the log odds of failure among individuals with $\boldsymbol{x}_{ij} = 0$, and $\boldsymbol{\beta}$ is a vector of known log odds ratios for failure. Now assume that the target concentrations has conditional distribution

$$f(y_{ij}|z_{ij}) = f_1(y_{ij})z_{ij} + f_2(y_{ij})(1 - z_{ij}) \tag{3.2}$$

where f_1 is the distribution of y for failures and f_2 is the distribution of y for non-failures.

Biological measurements are known to be imperfect measures of the true, underlying quantity in biological specimens. Inaccuracies can arise due to inadequate mixing prior to extracting test aliquots, random aliquot variability, or imprecise testing technology. In general it is not possible to test the same aliquot more than once, and hence it is not

possible to disentangle these sources of error. We therefore assume a single measurement error component that encompasses all sources of error and variation. We consider a general measurement-error function $y^* = h(y, \epsilon)$ where y^* is the observed quantity, y is the true, unobservable quantity, and ϵ is a random measurement-error component. The function h is deterministic, similar to a link function, while ϵ has density $g(\epsilon)$. Both h and g could take a variety of forms, for example, Normal(0,1) additive measurement error would be formulated using $h(y, \epsilon) = y + \epsilon$ and $g(\epsilon) = (2\pi)^{-\frac{1}{2}} exp(-\frac{1}{2}\epsilon^2)$. Given h and g the observed row and column-pool measurements can then be expressed as

$$r_i = h(\frac{1}{n}\sum_{j=1}^{n} y_{ij}, \epsilon_i) \tag{3.3}$$

$$c_j = h(\frac{1}{n}\sum_{i=1}^{n} y_{ij}, \epsilon_j) \tag{3.4}$$

where all ϵ_i and ϵ_j have density function g. Define $\boldsymbol{r} = \{r_i\}$, an $n \times 1$ vector of row-pool values, and $\boldsymbol{c} = \{c_j\}$, the $n \times 1$ vector of column pool values.

The goal is to identify individual cells most likely to be failures, conditional on the observed values $\boldsymbol{r}$, $\boldsymbol{c}$, and covariates $\boldsymbol{x}$. This can be thought of as finding the specific configuration of failures and non-failures most likely to have produced the observed data under the given distributional assumptions. That is, find

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}}{\text{argmax}}(f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{Z}, \Theta, \boldsymbol{x})) \tag{3.5}$$

where Θ is the collection of model parameters. In some cases the distribution $f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{Z}, \Theta, \boldsymbol{x})$ may be expressed in closed form, such as when y is a mixture of normals, ϵ is normally distributed and measurement error is additive (Hanscom et al., 2014a). However for important

examples (such as HIV viral load measurement) when y can reasonably be expected to follow a mixture of lognormal distributions, $f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{Z}, \Theta, \boldsymbol{x})$ does not have a closed form, and hence it is difficult maximize over $\boldsymbol{Z}$. The likelihood has the general form

$$L(\Theta; \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{x}) = \int_Y \int_Z f(\boldsymbol{r}|\boldsymbol{Y}, \tau) f(\boldsymbol{c}|\boldsymbol{Y}, \tau) f(\boldsymbol{Y}|\boldsymbol{Z}, \lambda, \theta, \sigma_f, \sigma_n) f(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\beta}). \quad (3.6)$$

Instead we consider a Bayesian approach and MCMC estimation to identify the most likely configuration of $\boldsymbol{Z}$, conditional on the observed data, and simultaneously estimate the posterior distribution of Θ. In general we will use normal priors for means and gamma priors for variances, with specific prior parameter values chosen depending on the application. We assume strong existing knowledge of the distributional parameters for both failures and non-failures, and hence apply informative priors with minimal variance. If the set of prior distributions for Θ is indicated by $\pi(\Theta)$, then the posterior will be proportional to the joint conditional distribution of all unknown quantities $\boldsymbol{Y}, \boldsymbol{Z}, \Theta$, written as

$$f(\boldsymbol{Y}, \boldsymbol{Z}, \Theta|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{x}) \propto f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{Y}, \boldsymbol{Z}, \Theta, \boldsymbol{x}) \quad (3.7)$$

$$= f(\boldsymbol{r}|\boldsymbol{Y}, \Theta) f(\boldsymbol{c}|\boldsymbol{Y}, \Theta) f(\boldsymbol{Y}|\boldsymbol{Z}, \Theta) f(\boldsymbol{Z}|\boldsymbol{x}, \Theta) \pi(\Theta). \quad (3.8)$$

In a typical problem we would be interested in the posterior distribution of Θ, namely $f(\Theta|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{x})$, but in this case our interest is primarily in the conditional distribution $f(\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{x})$. By constructing and MCMC chain based on (3.8) we can effectively sample from both of these distributions at the same time (Tanner and Wong, 1987). Provided that the MCMC sampler effectively explores the joint sample space, samples of any individual parameter or latent variable will represent a sample from its respective marginal distribution (marginal with respect to the other unknown quantities, but still conditional on observed data). The set of complete conditional distributions needed to construct the MCMC sequence is out-

lined in Appendix A.

Let $\boldsymbol{Z}^{(1)}...\boldsymbol{Z}^{(N)}$ be a set of sampled values of $\boldsymbol{Z}$ from N MCMC iterations, and let $\boldsymbol{Z}^*$ be the most frequently occurring failure configuration in that sample, i.e. the multivariate sample mode. Each individual specimen in the matrix corresponding to $z^*_{ij} = 1$ is then tested for viral load. This testing procedure yields some number of observed values of $\boldsymbol{Y}$, indicated as $\boldsymbol{y}^*$. The distribution of $\boldsymbol{Z}$ can then be written conditionally on $\boldsymbol{y}^*$ (as well as $\boldsymbol{r}$ and $\boldsymbol{c}$), where the measurement-error process is assumed to be the same as for $\boldsymbol{r}$ and $\boldsymbol{c}$ so that $y^*_{ij} = h(y_{ij}, \epsilon_{ij})$. Note that measurement error is assumed to be independent for observed quantities $\boldsymbol{y}^*$, $\boldsymbol{r}$, and $\boldsymbol{c}$ and therefore $\boldsymbol{y}^*$ is conditionally independent of $\boldsymbol{r}$ and $\boldsymbol{c}$ given the true values $\boldsymbol{Y}$. We can therefore decompose the distribution of all unknown quantities as

$$f(\boldsymbol{Z}, \boldsymbol{Y}, \Theta|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{x}) \propto f(\boldsymbol{r}|\boldsymbol{Y}, \tau) f(\boldsymbol{c}|\boldsymbol{Y}, \tau) f(\boldsymbol{y}^*|\boldsymbol{Y}, \tau) f(\boldsymbol{Y}|\boldsymbol{Z}, \lambda, \theta, \sigma) f(\boldsymbol{Z}|\boldsymbol{\beta}, \boldsymbol{x}) \pi(\Theta) \quad (3.9)$$

and sample $\boldsymbol{Z}$ conditionally on $\boldsymbol{r}$, $\boldsymbol{c}$, $\boldsymbol{x}$, and $\boldsymbol{y}^*$. MCMC sampling is again used to generate a new sample $\boldsymbol{Z}^{(1)}...\boldsymbol{Z}^{(N)}$, and the mode $\boldsymbol{Z}^*$ is again identified. For each element of $\boldsymbol{Z}^*$ such that $z^*_{ij} = 1$, the corresponding sample is tested individually provided that it has not been tested already. This process is repeated until the current value of $\boldsymbol{Z}^*$ does not identify any new candidates for individual testing, at which point the algorithm is complete.

Because $\boldsymbol{Z}$ is a matrix, and because we are interested in a particular $\boldsymbol{Z}^*$ that determines the next subsequent testing rule, we define "convergence" of the MCMC sequence in a non-standard way. Rather than assessing whether the mean or median of each element z_{ij} of $\boldsymbol{Z}$ is stable, we instead focus on whether the mode $\boldsymbol{Z}^*$ is stable. Once the point in the MCMC is reached where the mode does not change, we consider the sequence to have "converged", since for the purposes of the testing algorithm, further iterations would not change which

individual samples are tested next, and hence not change the results.

3.3.3 MBS for Viral Load Assessment

Biological concentration values such as viral load are often found to be lognormally distributed, and so for our application we let y_{ij} have the conditionally lognormal distribution

$$y_{ij}|z_{ij} \sim \text{LN}(\lambda z_{ij} + \theta(1 - z_{ij}), \sigma_f^2 z_{ij} + \sigma_n^2(1 - z_{ij})) \tag{3.10}$$

where λ is the mean log concentration among failures, θ is the mean log concentration among non-failures and σ_n and σ_f are the standard deviations of log concentration among non-failures and failures, respectively. The marginal distribution of y_{ij} is therefore a mixture which can be expressed as

$$f(y_{ij}|\lambda, \theta, \sigma_f^2, \sigma_n^2) = pf_{LN}(\lambda, \sigma_f^2) + (1 - p)f_{LN}(\theta, \sigma_n^2) \tag{3.11}$$

where $f_{LN}(.,.)$ is the lognormal density and p is the failure prevalence.

Measurement error is assumed to have constant variance on the log scale, and so we define as $h(y, \epsilon) = exp(log(y) + \epsilon)$ with $\epsilon \sim N(0, \tau^2)$. The observed row and column-pool measurements can then be expressed as

$$r_i = \exp(\ln(\frac{1}{n}\sum_{j=1}^{n} y_{ij}) + \epsilon_i) \tag{3.12}$$

$$c_j = \exp(\ln(\frac{1}{n}\sum_{i=1}^{n} y_{ij}) + \epsilon_j) \tag{3.13}$$

The joint distribution (3.8) is then

$$f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Y},\boldsymbol{Z},\Theta|\boldsymbol{x}) = f(\boldsymbol{r}|\boldsymbol{Y},\tau)f(\boldsymbol{c}|\boldsymbol{Y},\tau)f(\boldsymbol{Y}|\boldsymbol{Z},\lambda,\theta,\sigma_f,\sigma n)f(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\beta})\pi(\Theta) \tag{3.14}$$

where

$$f(\boldsymbol{r}|\boldsymbol{Y},\tau) = \prod_{i=1}^{n}(r_i\sqrt{2\pi}\tau)^{-1}e^{-(1/2\tau^2)(\ln(r_i)-\ln(\frac{1}{n}\sum_{j=1}^{n}y_{ij}))^2} \tag{3.15}$$

$$f(\boldsymbol{c}|\boldsymbol{Y},\tau) = \prod_{j=1}^{n}(c_j\sqrt{2\pi}\tau)^{-1}e^{-(1/2\tau^2)(\ln(c_j)-\ln(\frac{1}{n}\sum_{i=1}^{n}y_{ij}))^2} \tag{3.16}$$

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{Z},\lambda,\theta,\sigma_f,\sigma_n) &= \prod_{ij}(y_{ij}\sqrt{2\pi}\sigma_{ij})^{-1} \\ &\quad \times e^{-(1/2\sigma_{ij}^2)(\ln(y_{ij})-\ln(\lambda z_{ij}+\theta(1-z_{ij}))^2} \end{aligned} \tag{3.17}$$

$$f(\boldsymbol{Z}|\beta,\boldsymbol{x}) = \prod_{ij}(\text{expit}(\boldsymbol{x}_{ij}\boldsymbol{\beta}))^{z_{ij}}(1-\text{expit}(\boldsymbol{x}_{ij}\boldsymbol{\beta}))^{1-z_{ij}} \tag{3.18}$$

and $\Theta = (\lambda,\theta,\sigma_f,\sigma_n,\tau,\boldsymbol{\beta})$, and $\sigma_{ij} = \sigma_f^2 z_{ij} + \sigma_n^2(1-z_{ij})$. Once a first round of individual testing is complete, the joint distribution (3.14) expands to

$$\begin{aligned} f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{y}^*,\boldsymbol{Y},\boldsymbol{Z},\Theta|\boldsymbol{x}) &= f(\boldsymbol{r}|\boldsymbol{Y},\tau)f(\boldsymbol{c}|\boldsymbol{Y},\tau)f(\boldsymbol{y}^*|\boldsymbol{Y},\tau)f(\boldsymbol{Y}|\boldsymbol{Z},\lambda,\theta,\sigma_f,\sigma_n) \\ &\quad \times f(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\beta})\pi(\Theta) \end{aligned} \tag{3.19}$$

where $y_{ij}^*|y_{ij} \sim \text{LN}(y_{ij},\tau^2)$, giving

$$f(\boldsymbol{y}^*|\boldsymbol{Y},\tau) = \prod_{ij'}(y_{ij}^*\sqrt{2\pi}\tau)^{-1}e^{-(1/2\tau^2)(\ln(y_{ij}^*)-\ln(y_{ij}))^2} \tag{3.20}$$

where ij' indicate the coordinates of all specimens that have been tested individually. There will instances when viral loads will be so low as to be undetectable. In these instances y^*_{ij} is set to the approximate midpoint of the undetectable range.

3.3.4 Selecting priors

The method described here is in some ways opposite to typical model estimation where randomly sampled data is used to estimate model parameters. Here we use prior information about the parameters, combined with aggregated data (row and column pools), to estimate something about the unobserved data points, i.e. which $z_{ij} = 1$.

The use of informative priors is intended first and foremost to provide information helpful for identifying failures in a pooled matrix. A second function is to prevent pooled and individual test results from having too much influence on the posterior distributions of the parameters. Test results from a specific matrix provide critical information for pinpointing which cells in that particular matrix are failures, but in general they will amount to a very small sample from the larger population and hence provide minimal information regarding the overall population parameters. By specifying informative priors on the parameters, we effectively separate the information provided by strong outside knowledge (based thousands of laboratory samples, for example) from the information provided by the handful of pooled and individual tests an a matrix.

Information for prior distributions will depend on the number of samples available at a particular laboratory for the population of interest. If a large number of samples are available, for example 100 failures and 1000 non-failures, prior distributions for λ, θ, σ_f, and σ_n may be specified using the sample means, sample variances, and sample standard errors of the means and variances. Letting $y1_i$ denote the viral load measurement for a

failure, then the priors for the failure mean and variance would be

$$\pi(\lambda) = N\left(\frac{1}{100}\sum_{i=1}^{100}\ln(y1_i), \frac{1}{\sqrt{100}}S\right) \quad (3.21)$$

$$\pi(\sigma_f) = Gamma(\zeta_1, \zeta_2) \quad (3.22)$$

where

$$S = \sqrt{\frac{1}{99}\sum_{i=1}^{100}\left(\ln(y1_i) - \frac{1}{100}\sum_{i=1}^{100}\ln(y1_i)\right)^2} \quad (3.23)$$

and ζ_1 and ζ_2 are chosen such that $\zeta_1/\zeta_2 = S$ and $\zeta_1/\zeta_2^2 = (2/99)S^4$. Priors for non-failures could be chosen similarly by computing sample quantities based on the 1000 non-failure samples.

In the early stages of viral monitoring for a specific population, however, very few samples may be available and priors would need to be estimated based on expert knowledge from from other labs. In these cases prior variances could be larger, say two or three times the sample standard errors from outside lab data, to reflect the lack of specific knowledge regarding the lab/population of interest. This would allow pooled and individual data from an early matrix to have stronger influence on the posteriors and hence on the testing algorithm.

3.4 Simulation studies

To evaluate these methods we simulated random samples of n^2 specimens for a variety of different values of n and prevalence p. We compare the MBS method to the Simple Search, Modified Simple Search, and to individual testing, assessing turn-around time (testing rounds), efficiency, Sensitivity, Specificity, Positive Predictive Value(PPV), and Negative

Predictive Value (NPV). Efficiency is defined as the percent reduction in the number of tests used, as compared to testing each individual separately. We evaluate the MBS under two scenarios. In the first scenario we assume that no covariate data is available, and refer to this version as MBS-1. In the second scenario, MBS-2, we assume that covariate data is available and has a known, significant association with the probability of treatment failure.

This method can be used for any application where quantitative testing is performed. Although we use the term "failure" to indicate a concentration value above a critical threshold, in other settings a high concentration may indicate "success". The methodology is the same regardless of how a threshold breach is labelled.

3.4.1 HIV Viral Load Example - A Simulation Study

In resource-wealthy settings, HIV-infected individuals who are being treated with antiretroviral medications (ART) are routinely monitored for virologic failure, defined as a detectable proliferation of HIV virus in the blood despite treatment. Individuals experiencing virologic failure may decline into worse health and experience AIDS, and they also may present an increased risk of transmitting treatment-resistant virus to sexual partners. In the context of low-income countries, regular viral-load testing (performed by reverse transcriptase polymerase chain reaction (RT PCR)) is expensive and time consuming. Limited HIV funds must also be allocated to identifying HIV cases and providing treatment, education, prevention interventions, and a host of other services. In these settings it is therefore critical to minimize the cost of viral monitoring.

Empirical data suggest that viral-load values tend to be skewed and are often well described by the lognormal distribution. In addition, assay measurement error tends to be constant on the log scale, with assay standard deviations of approximately 0.12 on the log_{10}

scale (Brambilla et al., 1999), suggesting substantial variation for large viral loads.

Simulated viral load values for failure and non-failures were were generated as lognormal using the parameter values in Table 3.1. Failure values were shifted by +1000, giving an actual mean of about 11,000. Measurement error was generated on the natural log scale as normal with mean 0 and standard deviation 0.276 (equal to 0.12 on the log10 scale). The probability of failure was associated with a single, Normal(0,1) covariate x via equation (3.1), with relative risk $e^{\beta} = 5.0$.

Table 3.1: Parameter values used in the simulation study

Parameter	Value
λ	8.5
θ	4.0
σ_f	1.0
σ_n	0.50
τ	0.276
β	ln(5.0)

Two hundred datasets were generated for baseline prevalence levels ranging from 1% to 10%, and all four pooled-test methods were run on each dataset. The actual number of failures in each simulated dataset was random and depended on both the baseline prevalence and the values of x. Because the effect of x was large, the range of observed failure prevalences was much larger than the range of baseline prevalances (See Tables 3.3 and 3.4). Prior distributions for the model parameters were assumed to be as listed in Table 3.2. A first set of simulations were generated using 150,000 MCMC iterations per test round, and then a second set of simulations using 500,000 iterations per round. There was virtually no difference between the two sets of simulation results, and therefore 500,000 was considered sufficient to achieve convergence for the purposes of the MBS algorithm.

Table 3.2: Prior distributions and starting values for model parameters in the simulation study

Parameter	Prior Distribution	Starting Value
λ	Normal(8.5, 0.0025)	8.5
θ	Normal(4.0, 0.0025)	4.0
σ_f	Gamma(10000,10000)	1.0
σ_n	Gamma(5000,10000)	0.50
p	Uniform(0,1)	0.10

The relative performance of the various methods depends heavily on the underlying prevalence, with all methods performing fairly well at low prevalences, and the MSS and MBS preforming relatively better at higher prevalances (Tables 3.3 and 3.4). At 2% failure prevalence, the MSS and MBS methods reduce the number of rounds from about five to around two or three, and all three methods have high efficiency (about 75%) and nearly 100% sensitivity. At 10% prevalence the reduction in turn-around time (rounds) is more pronounced for MSS and MBS, with only about six rounds needed as compared to 24 rounds for SS. When the covariate information is used (MBS-2), the number of rounds required drops to just over four. In terms of efficiency SS and MBS-1 perform similarly well at about 57%, with MSS trailing at 53%, and MBS-2 showing a substantial advantage at 65%. The MBS methods show modestly better sensitivity than SS and MSS. A similar pattern holds at higher prevalances, with MSS and both MBS methods dramatically reducing turn-around time, only MBS-2 showing strong a efficiency advantage, and both MBS methods modestly improving sensitivity. Note that even individual testing does not achieve perfect sensitivity, and this is due to measurement error.

Also note that although exactly 100 datasets were generated using each baseline prevalence level 1% through Tables 3.3 and 3.4, the actual number of failures in each simulated dataset is random and depends on the value of the randomly generated covariate. This is why the numbers of simulated datasets for each prevalence level is not the same.

Table 3.3: Simulation results for 10x10 matrix pools Part 1, failure prevaluence 1%-10%. Lognormal data, with 500,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failure prevalence (%)									
	1	2	3	4	5	6	7	8	9	10
n	61	76	118	118	118	111	136	130	158	150
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	2.7	4.6	6.5	8.6	10.8	13.6	16.0	19.4	22.0	24.2
Rounds - Mod. Simple Search	2.0	2.1	2.6	3.1	3.5	4.0	4.3	4.9	5.5	5.9
Rounds - MBS-1	2.4	2.9	3.3	3.3	3.9	4.4	4.9	5.1	5.3	5.8
Rounds - MBS-2	2.3	2.5	2.7	2.7	3.1	3.3	3.4	3.7	3.9	4.3
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	78.3	76.4	74.5	72.4	70.2	67.4	65.0	61.6	59.0	56.8
Efficiency - Mod. Simple Search	77.9	74.6	72.4	69.9	67.3	64.7	62.1	58.8	55.6	53.0
Efficiency - MBS-1	77.9	76.0	74.1	72.9	70.3	67.8	65.4	63.1	60.7	57.5
Efficiency - MBS-2	78.3	76.8	75.4	74.5	72.9	71.6	70.3	68.6	67.1	65.0
Sensitivity - Individual Testing	100	100	100	99.4	99.8	99.5	99.2	100	99.5	99.5
Sensitivity - Simple Search	100	98.7	95.2	94.5	93.4	90.8	89.8	89.4	87.6	87.4
Sensitivity - Mod. Simple Search	100	100	99.4	97.9	95.4	93.2	92.0	91.5	89.2	88.7
Sensitivity - MBS-1	100	100	99.4	98.5	97.3	95.3	94.7	94.0	91.5	91.8
Sensitivity - MBS-2	100	100	99.7	98.3	97.8	96.4	95.5	94.8	93.5	93.1
NPV - Individual Testing	100	100	100	100	100	100	99.9	100	100	99.9
NPV - Simple Search	100	100	99.9	99.8	99.7	99.4	99.2	99.1	98.8	98.6
NPV - Mod. Simple Search	100	100	100	99.9	99.8	99.6	99.4	99.3	99.0	98.8
NPV - MBS-1	100	100	100	99.9	99.9	99.7	99.6	99.5	99.2	99.1
NPV - MBS-2	100	100	100	99.9	99.9	99.8	99.7	99.6	99.4	99.2

Table 3.4: Simulation results for 10x10 matrix pools Part 2, failure prevaluence 11%-20%. Lognormal data, with 500,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failure prevalence (%)									
	11	12	13	14	15	16	17	18	19	20
n	154	154	163	148	144	151	115	123	87	82
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	27.4	30.8	31.8	35.2	36.8	38.3	41.8	43.8	43.5	45.9
Rounds - Mod. Simple Search	6.5	7.1	7.4	7.9	8.3	8.7	9.1	9.4	9.6	9.8
Rounds - MBS-1	6.3	6.4	6.7	7.1	7.2	7.3	7.3	7.7	7.7	8.0
Rounds - MBS-2	4.3	4.6	4.9	5.1	5.0	5.4	5.2	5.7	5.3	5.6
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	53.6	50.2	49.2	45.8	44.2	42.7	39.2	37.2	37.5	35.1
Efficiency - Mod. Simple Search	49.4	45.6	43.8	40.7	38.6	36.5	34.3	32.0	30.9	29.2
Efficiency - MBS-1	53.3	50.8	48.3	45.2	42.6	39.7	38.6	33.7	32.4	29.6
Efficiency - MBS-2	63.9	61.8	60.0	58.3	57.1	54.3	53.9	50.3	50.9	47.4
Sensitivity - Individual Testing	99.7	99.6	99.7	99.5	99.6	99.7	99.6	99.8	99.5	99.7
Sensitivity - Simple Search	88.2	86.9	85.7	87.1	83.5	83.1	85.6	83.8	83.2	82.3
Sensitivity - Mod. Simple Search	89.1	88.7	87.5	88.2	85.0	84.8	85.5	84.9	84.5	84.8
Sensitivity - MBS-1	91.3	90.0	89.9	88.6	87.7	88.6	87.3	87.6	86.4	87.0
Sensitivity - MBS-2	91.6	90.7	90.5	89.6	88.2	89.2	87.4	88.0	87.1	87.2
NPV - Individual Testing	100	99.9	100	99.9	99.9	99.9	99.9	100	99.9	99.9
NPV - Simple Search	98.6	98.3	97.9	98.0	97.2	96.9	97.2	96.6	96.3	95.8
NPV - Mod. Simple Search	98.7	98.5	98.2	98.1	97.5	97.2	97.2	96.8	96.5	96.4
NPV - MBS-1	98.9	98.7	98.5	98.2	97.9	97.9	97.5	97.4	97.0	96.9
NPV - MBS-2	99.0	98.8	98.6	98.4	98.0	98.0	97.5	97.5	97.1	96.9

3.4.2 Sensitivity to assumed parameter values

The proposed method relies on the important assumption that we have strong prior information about the means and variances of the viral load distributions. In order to evaluate whether the proposed method is sensitive to incorrect specification of these parameters, we ran a number of additional simulation studies. In the first study, simulated viral-loads were generated using the same distributions as in the previous section, with non-failures and failures having mean viral loads of about 60 and 9,000 respectively. For both of the MBS algorithms we then assumed that the mean viral load among failures was 2,000 copies higher, or 11,000 copies per/mL, and ran both models with a corresponding strong prior on λ. Results are displayed in Tables 3.5 and 3.6. The relative performance of MBS-1 and MBS-2 compared to the Simple Search methods was essentially unchanged by the incorrect assumption regarding λ. As in the prior simulation, the performance of all methods was similar at low prevalences, and the MBS methods were superior at higher prevalences, particularly with respect to reducing testing rounds. For example, at 15.6% prevalence the Simple Search method required about 21 testing rounds, the Modified Simple Search about 7 testing rounds, MBS-1 5.6 testing rounds, and MBS-2 just under 4 testing rounds. Testing efficiency for MBS-2 was more than 10% higher than any other method, at 55% efficient, with the next best being Simple Search and MBS-1 at about 44%.

A similar set of sensitivity simulations were run with mean viral loads among both failures and non-failures incorrectly assumed to be too low, with mean for non-failures assumed to be 20 copies/ML and the mean for failures assumed to be 6,000 copies/mL. Again the performance of both MBS methods was generally better than the Simple Search methods at higher prevalence levels, particularly in terms of reducing testing rounds. (See Tables 3.7 and 3.8) At the lower prevalence levels, however, both MBS methods performed poorly compared to the Simple Search methods in terms of both testing rounds and efficiency. These results suggest that at higher prevalance levels the MBS methods are robust to mis-specification

of the viral load distributions. However, caution should be taken if the prevalence is low and there is uncertainly regarding how well the viral load distributions are known. (Note that these sensitivity analyses were run using 8x8 matrices and 150,000 MCMC iterations to reduce computational demands as compared to 10x10 matrices. To give a general sense for computation time requirements, ten Linux processors working simultaneously required approximately four weeks to generate the results in Tables 3.3 and 3.4.)

Table 3.5: Sensitivity analysis simulation results for 8x8 matrix pools Part 1, failure prevaluence 1%-15%. Mean viral load for failures are assumed to be 2000 copies/mL higher than the true value used to generate the data. Lognormal data, with 150,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failure prevalence %									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
n	47	64	83	106	139	109	142	132	153	122
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	2.2	3.8	5.5	7.5	9.1	11.7	13.8	16.0	19.0	20.8
Rounds - Mod. Simple Search	2.0	2.1	2.7	3.3	3.7	4.3	5.0	5.7	6.4	6.9
Rounds - MBS-1	2.1	2.3	2.7	3.2	3.6	3.9	4.4	4.9	5.1	5.6
Rounds - MBS-2	2.1	2.2	2.3	2.5	2.8	2.8	3.2	3.8	3.8	3.9
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	73.1	70.7	67.9	64.9	62.3	58.3	55.0	51.6	46.9	44.1
Efficiency - Mod. Simple Search	73.0	69.0	66.3	63.0	60.2	56.2	52.3	47.5	43.7	39.9
Efficiency - MBS-1	72.9	70.9	68.0	64.9	62.2	59.6	56.0	50.8	48.3	43.3
Efficiency - MBS-2	72.9	71.3	69.4	67.5	65.2	63.9	61.5	58.3	56.9	54.8
Sensitivity - Individual Testing	100.0	100.0	99.6	99.8	100.0	99.8	99.9	99.9	100.0	99.9
Sensitivity - Simple Search	100.0	96.1	95.2	93.6	89.1	88.5	89.2	86.6	87.3	86.1
Sensitivity - Mod. Simple Search	100.0	100.0	98.8	96.0	91.2	89.8	90.2	89.1	87.7	87.4
Sensitivity - MBS-1	100.0	100.0	99.6	98.6	95.4	93.4	91.6	91.7	88.6	89.2
Sensitivity - MBS-2	100.0	100.0	99.6	98.1	97.0	94.6	92.8	93.3	90.6	90.1
NPV - Individual Testing	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
NPV - Simple Search	100.0	99.9	99.8	99.6	99.1	98.8	98.7	98.1	98.0	97.5
NPV - Mod. Simple Search	100.0	100.0	99.9	99.7	99.3	99.0	98.8	98.5	98.1	97.7
NPV - MBS-1	100.0	100.0	100.0	99.9	99.6	99.3	99.0	98.8	98.2	98.1
NPV - MBS-2	100.0	100.0	100.0	99.9	99.7	99.5	99.1	99.1	98.5	98.2

Table 3.6: Sensitivity analysis simulation results for 8x8 matrix pools Part 2, failure prevaluence 17%-30%. Mean viral load for failures are assumed to be 2000 copies/mL higher than the true value used to generate the data. Lognormal data, with 150,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failures prevalence									
	17.2	18.8	20.3	21.9	23.4	25	26.6	28.1	29.7	31.2
n	149	124	118	110	79	79	55	32	21	13
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	22.2	24.5	26.3	27.0	29.7	30.9	32.5	33.7	35.1	39.1
Rounds - Mod. Simple Search	7.2	7.8	8.3	8.4	9.1	9.5	9.6	10.2	10.4	11.6
Rounds - MBS-1	5.7	6.1	6.3	6.5	6.7	6.6	6.8	6.9	6.8	7.1
Rounds - MBS-2	4.2	4.6	4.7	4.7	4.8	4.7	4.9	5.1	5.5	5.3
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	41.8	38.3	35.5	34.4	30.2	28.2	25.8	23.9	21.7	15.5
Efficiency - Mod. Simple Search	38.0	34.9	31.5	31.2	26.7	24.1	23.0	19.7	18.6	10.7
Efficiency - MBS-1	41.7	38.2	34.5	30.6	28.9	26.4	24.9	22.8	19.4	16.1
Efficiency - MBS-2	53.0	49.5	47.3	46.4	44.0	42.7	40.4	39.7	34.7	33.5
Sensitivity - Individual Testing	99.5	99.7	99.7	99.7	99.7	99.5	99.9	99.8	99.7	99.2
Sensitivity - Simple Search	84.4	83.8	84.6	82.7	82.8	82.0	81.3	80.6	78.9	82.7
Sensitivity - Mod. Simple Search	86.0	84.2	86.2	83.6	82.7	83.5	82.9	81.8	80.2	83.8
Sensitivity - MBS-1	85.9	85.4	85.5	84.8	84.1	81.3	82.8	81.6	82.5	85.4
Sensitivity - MBS-2	87.1	87.7	87.5	86.0	85.8	84.8	84.0	83.3	85.5	87.7
NPV - Individual Testing	99.9	99.9	99.9	99.9	99.9	99.8	100.0	99.9	99.9	99.7
NPV - Simple Search	96.9	96.5	96.3	95.4	95.1	94.4	93.8	93.0	92.0	92.8
NPV - Mod. Simple Search	97.2	96.5	96.7	95.7	95.1	94.9	94.3	93.5	92.5	93.3
NPV - MBS-1	97.2	96.8	96.5	96.0	95.4	94.2	94.2	93.4	93.2	93.8
NPV - MBS-2	97.4	97.3	97.0	96.3	95.9	95.3	94.6	94.0	94.3	94.8

Table 3.7: Sensitivity analysis simulation results for 8x8 matrix pools Part 1, failure prevaluence 1%-15%. Mean viral load for failures are assumed to be 2000 copies/mL less than the true value used to generate the data, and the mean for non-failures was assumed to be 20 instead of 60 copies/mL. Lognormal data, with 150,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failure prevalence %									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
n	51	81	79	99	108	122	125	125	114	108
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	2.1	3.4	5.3	7.5	9.9	11.9	14.6	16.5	18.6	21.0
Rounds - Mod. Simple Search	2.0	2.0	2.8	3.3	3.9	4.5	5.1	5.8	6.2	6.8
Rounds - MBS-1	4.3	4.4	4.8	5.1	4.9	5.2	5.6	5.6	6.0	5.9
Rounds - MBS-2	3.9	4.0	4.1	4.5	4.1	4.2	4.5	4.4	4.7	4.7
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	73.3	71.2	68.3	64.8	61.0	58.0	53.8	50.8	47.5	43.7
Efficiency - Mod. Simple Search	73.2	69.4	66.7	63.0	58.9	55.0	51.1	47.0	44.4	40.6
Efficiency - MBS-1	60.6	58.9	57.1	53.2	52.6	49.1	44.7	40.4	37.2	35.8
Efficiency - MBS-2	61.9	61.4	60.4	56.7	57.9	55.5	53.3	50.9	48.7	45.8
Sensitivity - Individual Testing	100.0	100.0	99.6	100.0	99.6	99.9	99.8	99.9	99.8	99.4
Sensitivity - Simple Search	100.0	96.3	93.7	94.9	93.0	90.4	89.4	88.7	86.9	86.4
Sensitivity - Mod. Simple Search	100.0	100.0	97.0	96.0	94.4	92.2	90.5	89.7	88.7	86.2
Sensitivity - MBS-1	100.0	100.0	99.6	99.2	98.3	97.0	95.4	94.7	93.9	93.3
Sensitivity - MBS-2	100.0	100.0	99.6	99.7	97.8	96.4	96.1	96.2	94.9	94.0
NPV - Individual Testing	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
NPV - Simple Search	100.0	99.9	99.7	99.7	99.4	99.0	98.7	98.4	97.9	97.6
NPV - Mod. Simple Search	100.0	100.0	99.9	99.7	99.5	99.2	98.9	98.6	98.2	97.6
NPV - MBS-1	100.0	100.0	100.0	100.0	99.9	99.7	99.5	99.3	99.0	98.8
NPV - MBS-2	100.0	100.0	100.0	100.0	99.8	99.6	99.5	99.5	99.2	98.9

Table 3.8: Sensitivity analysis simulation results for 8x8 matrix pools Part 2, failure prevaluence 17%-30%. Mean viral load for failures are assumed to be 2000 copies/mL less than the true value used to generate the data, and the mean for non-failures was assumed to be 20 instead of 60 copies/mL. Lognormal data, with 150,000 MCMC iterations per simulation, 50,000 of which were dropped for burn in. Continuous Normal(0,1) covariate with known association (odds ratio=5.0 per SD) with failure. MBS-2 uses covariate information, MBS-1 does not. Efficiency, Sensitivity, and NPV are percents. Specificity and Positive Predictive Values (PPV) are 100% in all cases.

	Failures prevalence									
	17.2	18.8	20.3	21.9	23.4	25	26.6	28.1	29.7	31.2
n	91	61	49	28	19	14	10	5	3	2
Rounds - Inidividual Testing	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Rounds - Simple Search	22.4	24.9	26.7	27.7	34.2	29.1	34.9	35.6	33.3	39.0
Rounds - Mod. Simple Search	7.2	7.9	8.3	8.6	10.1	9.4	10.5	10.6	11.0	11.0
Rounds - MBS-1	6.1	6.2	6.1	6.3	6.6	6.3	6.8	6.8	6.7	4.5
Rounds - MBS-2	4.9	5.1	4.8	5.5	5.4	5.2	5.7	5.2	6.3	4.0
Efficiency - Inidividual Testing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Efficiency - Simple Search	41.6	37.7	34.8	33.3	23.1	31.0	22.0	20.9	24.5	15.6
Efficiency - Mod. Simple Search	38.4	33.6	31.4	30.0	19.9	25.7	18.0	16.9	15.6	16.4
Efficiency - MBS-1	30.4	27.3	22.3	19.8	7.1	12.8	4.4	10.9	7.8	-4.7
Efficiency - MBS-2	43.5	40.2	39.4	33.2	30.3	29.7	21.4	28.4	27.1	15.6
Sensitivity - Individual Testing	99.4	100.0	100.0	100.0	99.6	99.6	98.8	100.0	100.0	100.0
Sensitivity - Simple Search	84.0	86.5	82.9	81.4	83.2	83.5	78.2	81.1	80.7	87.5
Sensitivity - Mod. Simple Search	85.0	86.6	84.0	82.4	83.9	83.5	78.8	82.2	82.5	85.0
Sensitivity - MBS-1	92.9	92.9	93.7	93.4	95.4	92.0	92.9	94.4	91.2	97.5
Sensitivity - MBS-2	93.9	94.1	90.6	95.2	94.7	91.5	94.7	94.4	94.7	95.0
NPV - Individual Testing	99.9	100.0	100.0	100.0	99.9	99.9	99.6	100.0	100.0	100.0
NPV - Simple Search	96.8	97.0	95.9	95.2	95.2	94.8	92.8	93.3	92.6	94.7
NPV - Mod. Simple Search	97.0	97.1	96.1	95.4	95.4	94.8	93.0	93.6	93.1	93.8
NPV - MBS-1	98.6	98.4	98.4	98.2	98.6	97.5	97.6	97.9	96.5	98.9
NPV - MBS-2	98.8	98.7	97.7	98.7	98.4	97.3	98.2	97.9	97.8	97.8

3.5 HIV Viral Load Examples

3.5.1 Matrix-pool data from a South African HIV population

To further test these methods we used viral-load data collected from the National Health Laboratory Service in Tygerberg, South Africa (van Zyl et al., 2011). Three sets of 10x10 matrix pools were available, with viral load measurements available for the pools as well as for the individual specimens. Unlike the simulation examples, we do not know the true viral load for any individual specimen y_{ij}, and so we let the failure indicator z_{ij} be equal to one if the observed value y^*_{ij} is larger than the failure threshold, and zero otherwise. For these analyses we set the virologic failure threshold to 1000 HIV RNA copies per mL. (The raw data is included in Appendix B.)

Also available in this data set were two covariates, time on ART (in months) and whether the ART regimen contained Nevirapine (NVP) or Efavirenz (EFV). A longer duration of viral suppression by ART has been shown to be strongly predictive of reduced risk of treatment failure (Rosenblum et al., 2009). Unfortunately about half of the treatment-duration data was incomplete, and so we classify each subject as either (a) treated for at least 12 months, (b) treated for less than 12 months, or (c) unknown. Based on the Rosenblum et al. (2009) results we assume that the odds ratio for failure for patients treated for 12 months or more is 0.75, as compared to patients treated for less than 12 months. We assume the patients with missing data comprise a mixture of patients with more or less than 12 months treatment, and let the odds ratio for failure for that group be 0.90. We also assume that treatment duration is a reasonable surrogate for viral suppression, which may be an optimistic assumption.

Pillay et al. (2013) conduct a meta analysis showing that, based on RCTs, patients taking an ART regimen that includes NVP are about 20% more likely to experience virologic failure than patients taking ART with EFV. For modeling purposes we assume an odds

ratio equal to 1.2 for the increased risk associated with NVP. It should be noted that there is no detectable marginal association between NVP/EFV and failure in this dataset, and so *a priori* we would not expect this covariate to dramatically improve testing efficiency with these samples.

For the MBS methods, prior distributions and starting values for the model parameters are shown in Table 3.9. Priors for λ, θ, σ_f, and σ_n are assumed to be fairly well estimated from existing lab data and hence the prior variances are very small. Starting values and prior means in this case come from the prior knowledge of field experts. The prevalence parameter p is designed to estimate the sample prevalence in each matrix, which is not known, and hence a minimally informative Uniform(0,1) prior is selected.

Table 3.9: Prior distributions and starting values for model parameters for HIV example data, South Africa)

Parameter	Prior Distribution	Starting Value
λ	Normal(10.6, 0.0025)	10.6
θ	Normal(3.8, 0.0025)	3.8
σ_f	Gamma(12828,10000)	1.28
σ_n	Gamma(5500,10000)	0.55
p	Uniform(0,1)	0.10

One and a half million MCMC iterations were tested for the MBS procedures, and little or no change occurred in $\boldsymbol{Z}^*$ after 1 million iterations. Therefore the MBS procedures used 1 million MCMC iterations at each testing round, with the first 100,000 dropped for burn in. Results are shown in Table 3.10. The relative performance of the three methods is somewhat inconsistent for the three matrices available here. For matrix 1, the MSS, MBS-1 and MBS-2 methods yield dramatically shorter turn-around times (fewer rounds), but show fairly similar efficiency as compared to SS, with MBS-1 having a slight advantage. Matrix 2 proved to be exceptionally challenging for all methods, and although MSS, MBS-1 and

MBS-2 required fewer testing rounds than SS, none of the methods provided substantial efficiency gains over individual testing. The MBS-1 method had a strong performance with matrix 3, requiring only half the individual tests used by SS and MSS, and just over half the testing rounds. In trade for this higher efficiency, MBS-1 missed just one failure caught by the other two methods (giving a sensitivity of 60% vs 70%). MBS-2 had a longer turn around time and worse efficiency than MBS-1, but it had the best sensitivity of all methods (80%). In summary, the MSS and the MBS methods always require fewer testing rounds than SS, and their efficiency tends to be similar or better than SS. It also appears that including covariates in the model is not always helpful and may hurt performance.

To give a sense for how much influence observed data has on parameter estimation, prior and posterior distributions for the first South Africa matrix are shown in Figure 3.3. Due to the very small number of failures sampled (seven), posteriors for failures are nearly identical to the priors. Relatively more samples (seventeen) from non-failures led to a prior-posterior shift for the non-failure distributions.

3.5.2 Matrix pool data from a United States HIV population

In a prior study of HIV patients in the US, 150 blood samples were collected from participants enrolled in the San Diego Acute Infection and Early Disease Research Program between January 1998 and January 2007 (Smith et al., 2009). The 150 specimens were divided into three 10x10 matrices by using each specimen twice (Matrix 1: samples 1-100, Matrix 2: samples 51-150, Matrix 3: samples 1-50, 101-150). For each matrix, row and column pools were formed, and all pools and individual specimens were tested. Viral load values above 500 copies/ml were considered failures. (The raw data is included in Appendix C.) No covariate information was available for these samples.

Table 3.10: Performance of three pooling methods for South African viral loads, 1,000,000 MCMC iterations. The number tested is the number of individual, non-pooled specimens tested.

	Tested	Rounds	Efficiency	Sensitivity	Specificity	PPV	NPV
Matrix 1							
Simple Search	27	27	0.53	1.00	1.00	1.00	1.00
Modified Simple Search	27	6	0.53	1.00	1.00	1.00	1.00
MBS-1	23	6	0.57	1.00	1.00	1.00	1.00
MBS-2	29	8	0.51	1.00	1.00	1.00	1.00
Matrix 2							
Simple Search	72	72	0.08	0.75	1.00	1.00	0.95
Modified Simple Search	68	13	0.12	0.75	1.00	1.00	0.95
MBS-1	75	11	0.05	0.81	1.00	1.00	0.97
MBS-2	67	11	0.13	0.75	1.00	1.00	0.95
Matrix 3							
Simple Search	33	33	0.47	0.70	1.00	1.00	0.97
Modified Simple Search	35	7	0.45	0.70	1.00	1.00	0.97
MBS-1	17	4	0.63	0.60	1.00	1.00	0.96
MBS-2	27	9	0.53	0.80	1.00	1.00	0.98

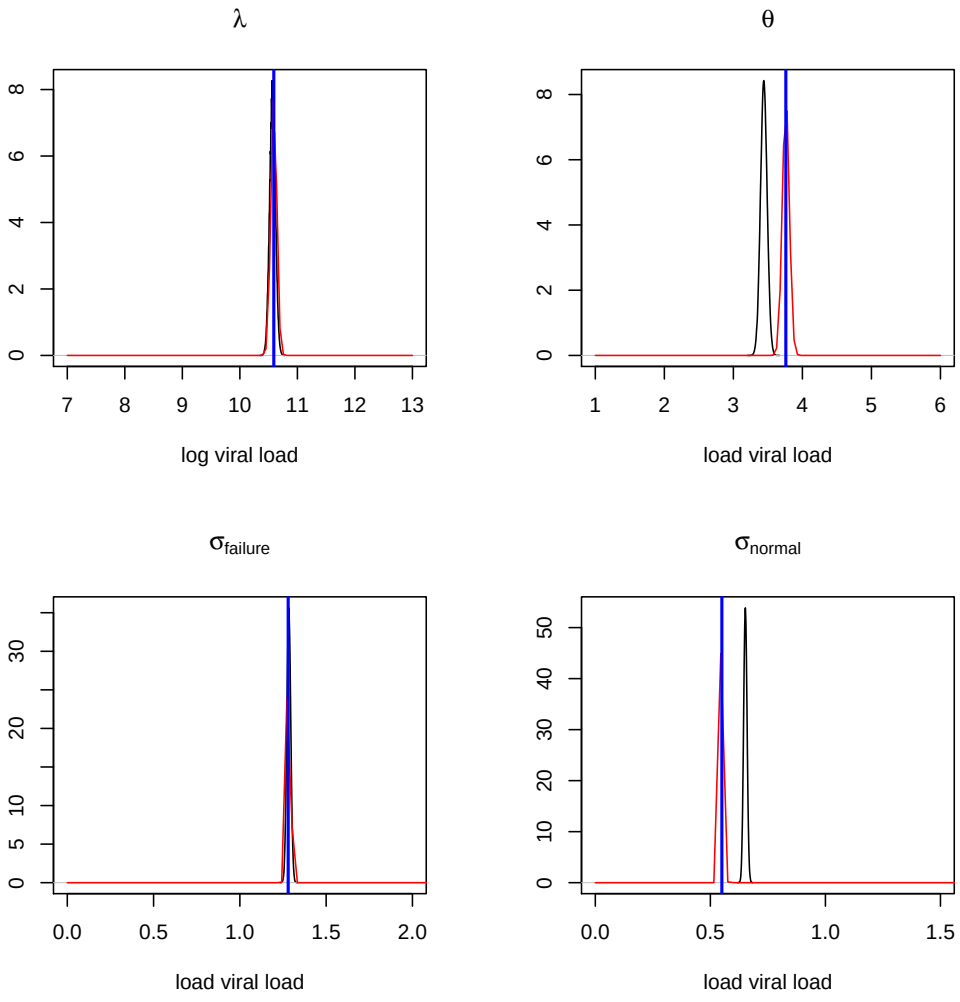

Figure 3.3: Prior distributions (red), starting values (blue) and posterior distributions (black) for the MBS-1 applied to the first South Africa pooled testing matrix.

For the MBS-1 method, prior distributions and starting values for the model parameters are shown in Table 3.11. As before, the values for λ, θ, σ_f, and σ_n are assumed to be fairly well estimated, and a minimally informative Uniform(0,1) prior is selected for p. The major difference in these data as compared to the South Africa data is that viral loads tend to me much lower, with a mean viral load of about 1,000 copies/ml among individuals experiencing treatment failure.

Table 3.11: Prior distributions and starting values for model parameters for HIV example data, San Diego, CA)

Parameter	Prior Distribution	Starting Value
λ	Normal(5.9, 0.0025)	5.9
θ	Normal(3.8, 0.0025)	3.8
σ_f	Gamma(8325,10000)	0.83
σ_n	Gamma(5500,10000)	0.55
p	Uniform(0,1)	0.10

We again ran 1.5 million MCMC iterations for all three matrices and found that matrix 2 and matrix 3 required only 250,000 iterations for convergence, but matrix 1 needed 1 million. For consistency the MBS-1 procedure used 1 million MCMC iterations for all cases, with the first 100,000 dropped for burn in. Results are shown in Table 3.12. The MBS1 and MSS methods strongly outperformed Simple Search in terms of turn-around time, requiring only 6-9 testing rounds as compared to 20-30 testing rounds. Results for testing efficiency, however, were not consistent, with SS showing better efficiency for matrix 1 and matrix 3, but worse efficiency for matrix 2. All three methods were identical in terms of sensitivity, performing well (100% sensitivity) on matrix 1 and matrix 2, less well on matrix 3 (67% sensitivity).

Table 3.12: Performance of three pooling methods for United States viral loads, 1,000,000 MCMC iterations. The number tested is the number of individual, non-pooled specimens tested.

	Tested	Rounds	Efficiency	Sensitivity	Specificity	PPV	NPV
Matrix 1							
Simple Search	33	33	0.47	1.00	1.00	1.00	1.00
Modified Simple Search	40	8	0.40	1.00	1.00	1.00	1.00
MBS-1	45	9	0.35	1.00	1.00	1.00	1.00
Matrix 2							
Simple Search	31	31	0.49	1.00	1.00	1.00	1.00
Modified Simple Search	25	6	0.55	1.00	1.00	1.00	1.00
MBS-1	25	8	0.55	1.00	1.00	1.00	1.00
Matrix 3							
Simple Search	21	21	0.59	0.67	1.00	1.00	0.99
Modified Simple Search	25	6	0.55	0.67	1.00	1.00	0.99
MBS-1	30	8	0.50	0.67	1.00	1.00	0.99

3.6 Discussion

Both new methods we propose here, the Modified Simple Search and the Model-Based Search performed well in simulation studies. Both can provide improved efficiency and dramatically reduced turn-around time as compared to the Simple Search Method proposed by May et al. (2010). If covariates are available that have strong, known associations with treatment failure, the benefit of using MBS becomes even larger. Unfortunately the advantage of using MSS and MBS was less definitive in our real-data examples than in simulation.

There are several possible reasons why matrix-pooling methods did not perform as well with the real viral load data examples. First, the data from South Africa appeared to include individual viral load data and pooled viral load data that seemed inconsistent. It was unfortunately not possible to retrospectively retest or otherwise validate these data. Second, the degree of measurement error and assay variation may be larger for these data/settings than previously reported for other data. Another possible source of measurement error might occur if samples are not well mixed. Then the aliquots used for the row pools may have different viral concentrations than the aliquots used for columns pools, and both of those may differ from the aliquot used for individual testing. Since we do not have repeat measurements for any given sample or pool, it is not possible to estimate either the measurement-error variance or the degree of aliquot variability.

A third issue relates to the fact that there can be enormous variation in the true serum viral load concentration among individuals experiencing treatment failure. Since measurement error tends to be lognormal, the error associated with a large viral measurement, say 100,000 copies per ml, could disguise a sample with a much smaller, but still failing, concentration of 5,000 that happened to fall into the same row or column pool. For example, in matrix 3 of the South Africa data (Appendix B) the third row pool has a measured concentration of 730 copies/ml. In the second column of that row the tested value is 20,000

copies/ml, which, if accurate, would imply that the row pool should contain at least 2,000 copies/ml. If this particular sample were tested first, both of the Simple Search methods would conclude that no more failures exist in row three, and completely miss the failure in column 9. The MBS methods would likely be thrown off by this large discrepancy as well, and sensitivity would likely suffer.

Conversely, if a column pool value were especially high, the algorithms might be lead to look for more failures than actually exist in that column, and efficiency would suffer. Consider the example of column one in matrix 2 (Appendix B) where the sample in row four measured 340,000 copies/ml, which would explain only 34,000 of the 240,000 copies/ml observed in the first column pool. All search methods were likely be led into searching for more failures in that column even though none exist.

An important strength of the MBS method is its flexibility in terms of distributional assumptions. As a direct result of using MCMC methods, nearly any distribution can be used to describe the failure and non-failure measurements. In addition, the measurement error function is flexible enough to handle a broad range of possible measurement-error patterns. This flexibility should allow the MBS method to be used in a variety of applications, should the need arise. The MSS method, on the other hand, has the advantage of being simple. No statistical analysis is necessary, and provided that strong covariate information is not available, the MSS may perform just as well as the MBS in many situations.

A limitation to the MBS methods is the requirement that most of the distributional parameters are known or well estimated. Although it would be nice to estimate or update these parameters based on the pooled data itself, it is unlikely that the very sparse data associated with matrix-pooled measurements would provide good information. However, in a laboratory setting where large numbers of biological samples are processed each day, it

would be quite reasonable to estimate distributional parameters for populations that are sampled regularly over time.

The strength of prior knowledge regarding viral load parameters will likely depend on the age of the laboratory and the duration of time that a particular population has been sampled and tested. In a situation where a lab was fairly new or where a new HIV population was being sampled, less informative prior distributions could be selected and the newly sampled data in a matrix allowed more influence. After each batch of tests, viral load values for individual samples could be incorporated into that lab's working parameter estimates, which would then serve as the prior estimates for next pooled testing matrix. A natural way to do this might be to use the posterior distributions from an MBS model as the prior distribution for the model in the next set of tests. This would incorporate information from both individual tests and row/columns pools into the priors for subsequent rounds. Another choice is to simply recompute parameters for prior distributions by combining previous and current individual test data and then using (3.21) and (3.22).

Further investigation should include an assessment of model performance under varying degrees of overlap between the failure distribution and non-failure distribution. The simulated examples reported here use distributions that have little overlap. Incremental increases in the amount of overlap between the normal and failure distributions will undoubtedly diminish the sensitivity and efficiency of all pooled testing methods.

In conclusion, both new methods we propose here - the MSS and MBS - can provide a fast, efficient, and sensitive testing algorithm, even in situations where failure prevalence is high. Unless covariates are available that have known associations with failure, it may be preferable to use MSS for the sake of simplicity. Implementation of these methods in the context of viral-load monitoring in resource-limited settings could yield substantial cost

savings as compared to individual testing methods.

3.7 Appendix A - Conditional distributions used in MCMC sampling.

When only the row and column pool data are available ($\boldsymbol{r}$ and $\boldsymbol{c}$) the MCMC sequence is generated using the following conditional distributions

$$\begin{aligned}
f(y_{ij}^{(n)}|z_{ij}^{(n-1)}, \boldsymbol{y}_{i.}^{(n-1)}, \boldsymbol{y}_{.j}^{(n-1)}, r_i, c_j, \Theta^{(n-1)}) &\propto f(y_{ij}|z_{ij}^{(n-1)}, \lambda^{(n-1)}, \theta^{(n-1)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) \\
&\quad \times f(r_i|\boldsymbol{y}_{i.}^{(n-1)}, \tau^{(n-1)}) \\
&\quad \times f(c_j|\boldsymbol{y}_{.j}^{(n-1)}, \tau^{(n-1)}) \qquad (3.24) \\
f(z_{ij}^{(n)}|y_{ij}^{(n)}, \Theta^{(n-1)}) &\propto f(y_{ij}^{(n)}|z_{ij}^{(n)}, \lambda^{(n-1)}, \theta^{(n-1)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) \\
&\quad \times f(z_{ij}^{(n)}|p^{(n-1)}) \qquad (3.25) \\
f(\tau^{(n)}|r_i, c_i, y_{ij}) &\propto \prod_{i=1}^{n} f(r_i|\boldsymbol{y}_{i.}^{(n)}, \tau^{(n)}) \prod_{j=1}^{n} f(c_j|\boldsymbol{y}_{.j}^{(n)}, \tau^{(n)}) \\
&\quad \times \pi(\tau^{(n)}) \qquad (3.26) \\
f(\lambda^{(n)}|y_{ij}^{(n)}, z_{ij}^{(n)}, \theta^{(n-1)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} f(y_{ij}^{(n)}|z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n-1)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) \\
&\quad \times \pi(\lambda^{(n)}) \qquad (3.27) \\
f(\theta^{(n)}|y_{ij}^{(n)}, z_{ij}^{(n)}, \lambda^{(n)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} f(y_{ij}^{(n)}|z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) \\
&\quad \times \pi(\theta^{(n)}) \qquad (3.28) \\
f(\sigma_f^{(n)}|y_{ij}^{(n)}, z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n)}, \sigma_n^{(n-1)}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} f(y_{ij}^{(n)}|z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n)}, \sigma_f^{(n)}, \sigma_n^{(n-1)}) \\
&\quad \times \pi(\sigma_f^{(n)}) \qquad (3.29) \\
f(\sigma_n^{(n)}|y_{ij}^{(n)}, z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n)}, \sigma_f^{(n)}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} f(y_{ij}^{(n)}|z_{ij}^{(n)}, \lambda^{(n)}, \theta^{(n)}, \sigma_f^{(n)}, \sigma_n^{(n)}) \\
&\quad \times \pi(\sigma_n^{(n)}) \qquad (3.30) \\
f(\beta_0^{(n)}|\boldsymbol{Z}^{(n)}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} f(z_{ij}^{(n)}|\beta_0^{(n)}, \boldsymbol{\beta}, \boldsymbol{x}) f(\beta_0^{(n)}) \qquad (3.31)
\end{aligned}$$

where the superscript (n) indicates the current MCMC iteration, the superscript $(n-1)$ indicates the prior MCMC iteration, and $\boldsymbol{y}_{i.}^{(n-1)}$ and $\boldsymbol{y}_{.j}^{(n-1)}$ are vectors whose elements take

on the values from $\boldsymbol{Y}^{(n)}$ for values previously sampled during the $(n)th$ iteration and take on values from $\boldsymbol{Y}^{(n-1)}$ for values not yet sampled during the $(n)th$ iteration.

Once individual samples have been tested, for any cell that has been tested the conditional distribution of $y_{ij}^{(n)}$ in (3.24) expands to

$$\begin{aligned} f(y_{ij}^{(n)}|z_{ij}^{(n-1)}, y_{ij}^{*}, \boldsymbol{y}_{i.}^{(n-1)}, \boldsymbol{y}_{.j}^{(n-1)}, r_i, c_j, \Theta^{(n-1)}) \quad &\propto \quad f(y_{ij}|z_{ij}^{(n-1)}, \lambda^{(n-1)}, \theta^{(n-1)}, \sigma_f^{(n-1)}, \sigma_n^{(n-1)}) \\ &\quad \times f(y_{ij}^{*}|y_{ij}^{(n)}, \tau^{(n)}) \\ &\quad \times f(r_i|\boldsymbol{y}_{i.}^{(n-1)}, \tau^{(n-1)}) \\ &\quad \times f(c_j|\boldsymbol{y}_{.j}^{(n-1)}, \tau^{(n-1)}). \end{aligned} \tag{3.32}$$

and (3.26) expands to

$$\begin{aligned} f(\tau^{(n)}|r_i, c_i, y_{ij}) \quad &\propto \quad \prod_{i=1}^{n} f(r_i|\boldsymbol{y}_{i.}^{(n)}, \tau^{(n)}) \prod_{j=1}^{n} f(c_j|\boldsymbol{y}_{.j}^{(n)}, \tau^{(n)}) \\ &\quad \times f(y_{ij}^{*}|y_{ij}^{(n)}, \tau^{(n)}) \times \pi(\tau^{(n)}) \end{aligned} \tag{3.33}$$

3.8 Appendix B

Table 3.13: Matrix 1 - Pooled viral load data from Tygerberg, South Africa (van Zyl et al., 2011). All quantities are HIV RNA copies per mL, with 0 indicating a value below the limit of detection.

	1	2	3	4	5	6	7	8	9	10	row pool
Matrix 1											
1	130	0	0	0	41	0	0	0	0	0	10
2	4700	0	0	0	0	2200	0	0	0	0	2800
3	0	0	0	0	0	0	0	0	0	0	10
4	0	0	0	0	0	40000	0	3500	26	30	46000
5	0	0	0	0	0	2100	0	0	0	0	1800
6	0	0	540	0	17000	0	0	0	0	1100	11000
7	0	0	0	0	0	0	0	0	0	0	10
8	0	0	0	0	0	110	0	0	0	0	10
9	98	380	0	0	0	0	0	0	0	48	160
10	0	39	120	0	31	0	94	0	0	0	10
column pool	3500	90	1100	10	8500	18000	10	190	10	400	

Table 3.14: Matrix 2 - Pooled viral load data from Tygerberg, South Africa (van Zyl et al., 2011). All quantities are HIV RNA copies per mL, with 0 indicating a value below the limit of detection.

	1	2	3	4	5	6	7	8	9	10	row pool
1	130	1200	2000	37	4000	0	0	50	470000	140	120000
2	0	95	1200	250	260	66	0	490	86	47	210
3	210	720	78	55	25	520	17000	190000	0	1100	17000
4	340000	0	190	0	0	40	0	0	0	0	160000
5	0	670	0	0	620	0	8000	0	0	0	1200
6	38	0	0	0	0	0	0	0	280	45	600
7	130	110	27	0	42	3700	34	3900	0	0	1300
8	0	430	0	150	0	0	1800	0	0	0	460
9	0	0	0	120	220	0	0	0	9600	6500	1600
10	0	0	0	200	1600	0	0	0	0	0	550
column pool	240000	350	1300	250	47	340	1400	44000	140000	2200	

Table 3.15: Matrix 3 - Pooled viral load data from Tygerberg, South Africa (van Zyl et al., 2011). All quantities are HIV RNA copies per mL, with 0 indicating a value below the limit of detection.

	1	2	3	4	5	6	7	8	9	10	row pool
1	0	37	0	0	97	120	50	0	0	3200	98
2	0	0	0	0	0	0	0	0	94	0	73
3	0	0	0	0	0	0	0	0	50	0	10
4	83	20000	0	0	0	0	0	0	6100	280	730
5	0	0	0	0	8200	0	0	0	0	0	2800
6	0	110	1800	0	0	0	0	0	0	0	71
7	0	0	0	0	0	2600	260	0	0	0	720
8	250000	560	0	230	430	64	41	65	0	170	31000
9	0	0	0	0	0	110	330	39	21000	0	17000
10	0	7700	0	0	110	2300	0	0	47	0	1600
column pool	26000	8000	10	54	2800	640	110	51	7400	150	

3.9 Appendix C

Table 3.16: Matrix pooled viral load data from San Diego, United States (Smith et al., 2009). All quantities are HIV RNA copies per mL, with 0 indicating a value below the limit of detection.

	1	2	3	4	5	6	7	8	9	10	row pool
Matrix 1											
1	0	0	193	53	51	0	0	408	0	0	65
2	0	183	0	58	2610	0	0	0	106	168	305
3	271	0	0	0	102	0	0	56	67	0	126
4	0	0	0	0	0	0	0	0	0	0	0
5	5180	0	0	0	0	0	0	0	0	0	505
6	0	0	0	0	0	0	0	0	0	445	72
7	419	371	0	0	51	0	0	0	0	0	80
8	0	0	0	0	0	0	105	0	96	0	0
9	0	0	130	246	0	474	0	892	362	0	324
10	488	112	0	0	0	0	0	0	0	0	66
column pool	796	103	92	0	143	195	0	482	238	51	
Matrix 2											
1	0	0	0	0	0	0	0	0	0	445	72
2	419	371	0	0	51	0	0	0	0	0	80
3	0	0	0	0	0	0	105	0	96	0	0
4	0	0	130	246	0	474	0	892	362	0	324
5	488	112	0	0	0	0	0	0	0	0	66
6	0	0	1390	0	0	0	0	0	0	0	138
7	0	0	0	0	0	0	0	0	55	0	0
8	0	0	0	120	83	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	375	0	91
10	0	0	0	0	0	0	0	473	0	0	86
column pool	62	98	199	136	0	100	0	275	264	0	
Matrix 3											
1	0	0	1390	0	0	0	0	0	0	0	138
2	0	0	0	0	0	0	0	0	55	0	0
3	0	0	0	120	83	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	375	0	91
5	0	0	0	0	0	0	0	473	0	0	86
6	0	0	193	53	51	0	0	408	0	0	65
7	0	183	0	58	2610	0	0	0	106	168	305
8	271	0	0	0	102	0	0	56	67	0	126
9	0	0	0	0	0	0	0	0	0	0	0
10	5180	0	0	0	0	0	0	0	0	0	505
column pool	504	0	0	114	148	0	0	154	186	88	

Chapter 4

AVOIDING MEASUREMENT-ERROR BIAS WHEN ESTIMATING PER COITAL ACT INFECTIVITY IN DISCORDANT PARTNER STUDIES

4.1 Abstract

Discordant partner studies are commonly used to quantify per-act infectivity rates for HIV. Statistical models used to estimate these rates depend on self-reported sexual activity, which is notoriously unreliable. The degree to which misreported sexual activity can affect infectivity estimates has not previously been reported. By using a basic transmission model in the context of measurement error, we show that infectivity estimates can be severely biased, and that the size and direction of bias depends on the underlying infectivity rate and on the mean number of sex acts occurring in the observation window. We show that by modifying the size of this window, in certain circumstances it is possible to avoid measurement-error bias without having to use a more sophisticated statistical model. We also show that, when misreporting is ignored, covariate parameter estimates can be biased, particularly when subgroup differences are large or when sexual frequency is heterogeneous.

4.2 Introduction

Each sexual contact between an individual infected with a sexually transmitted infection (STI) and an uninfected individual may result in transmission. The likelihood of such a transmission is referred to alternately as the "per-act infectivity" or "per-act transmission probability", and this quantity is of central interest in the study of STI prevention and epidemic modeling. The infectivity rate is needed for modeling infection dynamics, identifying infection risk factors, and designing prevention strategies. Also, differences in per-act in-

fectivity can serve as a primary measure of success or failure of STI-prevention interventions.

Prospective discordant-partner studies are one way to study disease transmission rates in a relatively controlled environment. In a typical discordant-partner study, one or both members of a partnership are interviewed at regular intervals regarding their sexual activity. Simultaneous to the interview, the disease-negative partner is tested to determine whether transmission has occurred. By counting the number of sexual contacts between each pair of partners prior to transmission, it is possible to estimate the per-act infectivity using the simple transmission model proposed by Jewell and Shibowski (1990).

This model of per-act infectivity must rely on self-reported sexual activity. However, it has been shown that retrospective reports of the number of sexual encounters are often inaccurate (Clark and Wallin, 1964). It also has been shown that the longer the time span over which people are asked to quantify their sexual activity, the less reliable those reports become (Kauth et al., 1991). Thus, even in the relatively controlled setting of prospective longitudinal discordant-partner studies, accurately quantifying sexual activity is a challenge.

It is well known that covariate measurement error can lead to severely biased regression parameter estimates (Fuller, 1987). In the case of linear models the amount and direction of bias can be easily estimated and corrected if the error distribution is known. But in more complicated non-linear models, measurement-error bias can be unpredictable and difficult to correct, particularly when replicate measurements do not exist and when the error distribution is unknown (Carroll et al., 2006). Prior infectivity studies have not addressed this issue and therefore the extent to which our current estimates of per-act infectivity suffer from measurement-error bias remains unknown.

In this paper we quantify the degree to which misreported sex acts can, when ignored,

induce biased estimates of per-act infectivity and related covariate parameters. We explore key variables that determine the amount and direction of bias, and show how modified partner-study designs can reduce or eliminate the potential for bias.

4.3 Modeling per act infectivity

A basic infectivity model was described by Jewell and Shibowski (1990) in the context of HIV partner studies. This model is derived from a typical time-to-event framework, but uses the number of sexual acts n in place of a measure of time, thus allowing for estimation of the per-act infectivity rate λ. Letting y be 1 if a transmission occurred after n sexual contacts, and 0 otherwise, the probability of transmission can be expressed as

$$P(y = 1) = 1 - (1 - \lambda)^n \tag{4.1}$$

The model assumes constant infectivity over time, and can be expanded to accommodate covariates X as follows,

$$P(y = 1) = 1 - (1 - \lambda)^{ne^{X\beta}}. \tag{4.2}$$

When λ is small equation (4.2) is approximately equal to

$$1 - (1 - \lambda e^{X\beta})^n. \tag{4.3}$$

and therefore β is approximately the log relative risk of per-act transmission per unit change in X. Estimation can be achieved using standard GLM optimization under the complementary log-log link function (Jewell and Shibowski, 1990).

This model and various extensions have been widely used in HIV partner studies (Hughes et al., 2012; Gray et al., 2001; Baeten et al., 2005). Although it has been used less frequently with other STIs where partner studies are less common, the model can be adapted to any setting where the infection status of each partner is known. Regardless of the application, misreporting of sex acts n is likely to be an issue.

4.4 Infectivity bias when misreporting is ignored

Under a simple scenario where the probability of disease transmission is as defined in equation (4.1) and the true number of sex acts n_i for couple i is distributed Poisson(ψ), we can compute the expected bias in $\hat{\lambda}$ for various values of ψ and λ. Assuming non-differential, Poisson measurement error for the reported number of acts m_i from a single partner, let

$$m_i \sim \text{Poisson}(n_i^*) \tag{4.4}$$

where $n_i^* = 0.25$ if $n_i = 0$ and $n_i^* = n_i$ otherwise. (This gives a median observed number of acts equal to the true median number of acts, and eliminates the possibility of defining a Poisson distribution with mean zero.) We assume that y_i and m_i are conditionally independent given n_i. If the m_i were assumed to be accurate, $\hat{\lambda}$ would typically be computed by maximizing the log likelihood given by

$$\log(f(\boldsymbol{y}|\lambda, \boldsymbol{m})) = \sum_{i=1}^{N} \log[\{1-(1-\lambda)^{m_i}\}^{y_i}\{(1-\lambda)^{m_i}\}^{1-y_i}] \tag{4.5}$$

where $\boldsymbol{y}$ and $\boldsymbol{m}$ are $(N \times 1)$ vectors containing the observed transmission indicators and act counts, respectively, for N couples. For any combination of λ and ψ the expected value of $\hat{\lambda}$ can be computed by finding the expected value of the first derivative of the log likelihood

Table 4.1: Expected percent bias for $\hat{\lambda}$ depending on the true infectivity λ and the true mean number of sex acts ψ, assuming Poisson measurement error.

	ψ											
λ	1	2	3	4	5	6	7	8	9	10	15	20
0.0001	-26	-12	-6	-3	-2	-1	0	0	0	0	0	0
0.0005	-26	-12	-6	-3	-2	-1	0	0	0	0	0	0
0.001	-26	-12	-6	-3	-2	-1	0	0	0	0	0	0
0.005	-26	-12	-6	-3	-2	-1	-1	0	0	0	0	0
0.01	-26	-12	-6	-3	-2	-1	-1	0	0	0	0	0
0.05	-26	-12	-6	-3	-2	-1	0	0	0	0	0	0
0.10	-27	-12	-6	-3	-1	0	0	0	1	1	1	2
0.15	-27	-12	-6	-3	-1	0	1	1	2	2	3	3
0.20	-27	-12	-6	-2	0	1	2	2	3	3	4	6
0.25	-27	-12	-5	-2	0	2	3	4	4	5	7	8
0.30	-27	-12	-5	-1	1	3	4	5	6	7	9	11
0.35	-27	-12	-5	-1	2	4	6	7	8	9	11	13
0.40	-27	-12	-4	0	3	5	7	8	10	11	14	16
0.45	-27	-11	-4	1	4	7	9	10	11	13	17	19
0.50	-27	-11	-3	1	5	8	10	12	13	15	20	23

function, setting it to 0, and solving for λ. (See Appendix A for details.) Numerical integration and maximization are required for these two steps. Expected percent bias is then computed as $(\hat{\lambda}-\lambda)/\lambda \times 100$, and its values for a variety of ψ and λ are presented in Table 4.1.

Several things stand out in the results shown in Table 4.1. First, depending on the values of λ and ψ in this example, the expected bias due to measurement error can be either positive or negative, with estimates ranging from 27% downward bias to 23% upward bias. Although it is unusual for measurement error to cause upward bias, it is not unheard of in non-linear models (Carroll et al., 2006). Second, strong downward bias exists for small values of ψ, and this bias is consistent for all values of λ. Third, there are a number of values of λ and ψ for which there is no bias, on average, particularly for scenarios with low λ and high ψ. This third point suggests that estimation bias might be partially controlled by study design, as discussed in Section 4.7.

The expected bias observed in Table 4.1 of course depends on the misreporting distribution as defined in (4.4). If the true measurement error variance is larger than the Poisson distribution allows, a negative binomial model might be more appropriate. For example, let

$$m_i \sim \text{NegBinomial}(\text{mean} = n_i^*, \text{size} = 10) \tag{4.6}$$

where, as before, $n_i^* = 0.25$ if $n_i = 0$ and $n_i^* = n_i$ otherwise. The *size* parameter in this example is chosen so that the negative-binomial variance is 50% larger than the Poisson variance when $n_i^* = 5$. As shown in Table 4.2, the increased measurement-error variance leads to more substantial bias at both the low and high ends of the parameter range. Larger measurement error also reduces the range of parameter values where no bias is expected. For example, when λ is between 0.0001 and 0.01, it is not until ψ is 9 or larger that the expected bias is zero. In the case of Poisson measurement error, ψ could be as low as 7 with no bias expected.

It should be noted that computation of *expected* bias for $\hat{\lambda}$ effectively assumes an infinite sample size. Finite sample statistics obtained in practice do not necessarily conform to this limiting case, as we describe in the next section.

4.5 Finite sample bias

Regardless of whether measurement error is present, in small-sample scenarios we expect a certain amount of estimation bias in non-linear modeling (McCulllagh and Nelder, 1983). When relatively small samples are combined with measurement-error, bias can be exacerbated. Here we conduct a simulation study to explore this issue in the case of the complementary log-log model.

Table 4.2: Expected percent bias for $\hat{\lambda}$ depending on the true infectivity λ and the true mean number of sex acts ψ, assuming negative binomial measurement error.

	ψ											
λ	1	2	3	4	5	6	7	8	9	10	15	20
0.0001	-28	-14	-8	-4	-3	-2	-1	-1	0	0	0	0
0.0005	-28	-14	-8	-4	-3	-1	-1	0	0	0	0	0
0.001	-28	-14	-8	-4	-3	-1	-1	0	0	0	0	0
0.005	-28	-14	-8	-4	-3	-2	-1	-1	0	0	0	0
0.01	-28	-14	-8	-4	-3	-2	-1	-1	0	0	0	0
0.05	-28	-14	-8	-4	-2	-1	-1	0	0	0	1	1
0.10	-28	-14	-7	-4	-2	-1	0	1	1	2	3	5
0.15	-28	-14	-7	-4	-1	0	1	2	3	4	7	10
0.20	-28	-14	-7	-3	0	1	3	4	5	6	11	16
0.25	-28	-14	-7	-2	1	3	5	6	8	9	16	22
0.30	-28	-13	-6	-1	2	5	7	9	11	12	21	29
0.35	-28	-13	-5	0	3	6	9	11	14	16	26	36
0.40	-28	-13	-5	1	5	8	11	14	16	19	31	43
0.45	-28	-12	-4	2	6	10	13	16	19	22	35	49
0.50	-28	-12	-4	2	7	11	15	18	22	25	39	54

For a range of pre-specified λ and ψ, 25,000 datasets were generated according to model (4.1) and measurement-error model (4.4), with sample sizes ranging from 500 to 5,000. For each simulated dataset, $\hat{\lambda}$ was estimated using GLM, and expected percent bias was estimated by computing the mean over all 25,000 simulations. Figure 4.1 displays results for six scenarios. When $\psi = 1$ we observe the negative 26% bias caused by measurement error, as predicted by Table 4.1, and small sample sizes do not yield any additional bias. Similarly, when $\psi = 5$ and when $\psi = 20$ along with $\lambda = 0.001$, little bias is present and sample size makes little or no difference. However, when $\psi = 20$ and $\lambda = 0.40$ the bias is substantially higher than what we expect due to measurement error alone, even for large samples. For datasets of size 500 the expected bias is 125%, as compared to the 16% expected based on measurement error alone. With samples of size 5,000 that bias is reduced to about 40%.

In general, when λ and ψ are small, finite samples do not add any additional bias above

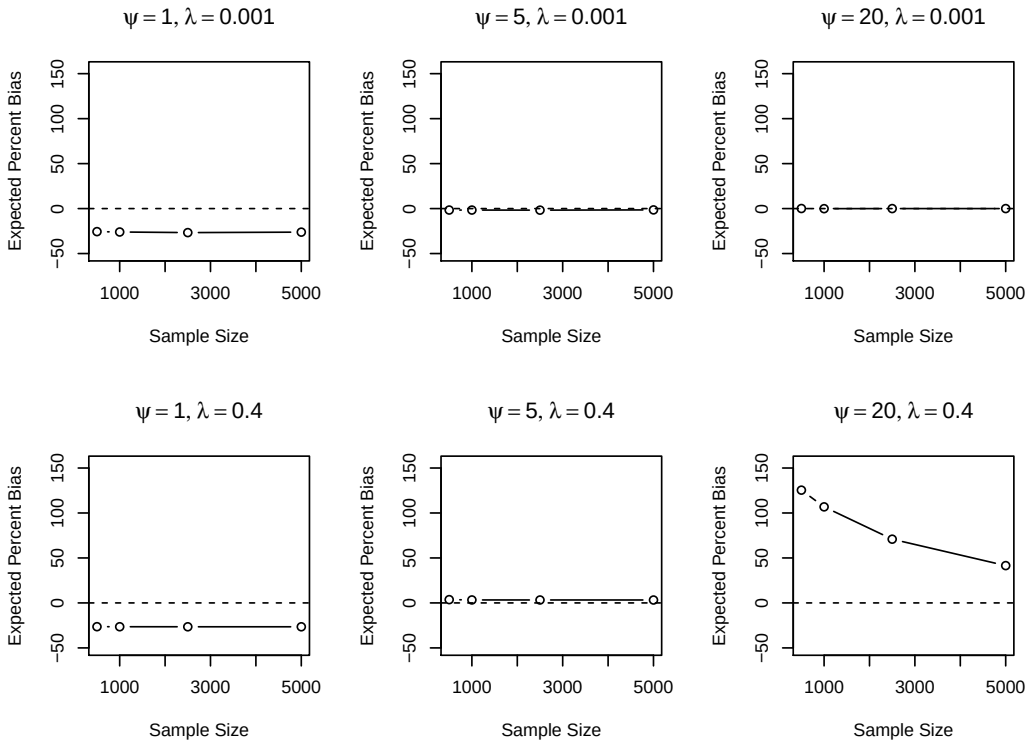

Figure 4.1: Expected percent bias for $\hat{\lambda}$ by sample size, for various combinations of λ (per-act infectivity) and ψ (mean number of contacts per couple).

Table 4.3: Simulation results for a covariate model showing how both infectivity and covariate parameter estimates are biased as a result of ignored measurement error. Each estimate is the mean over 200 simulated datasets.

							Percent Bias			
ψ	λ_0	β	$\lambda_1 = \lambda_0 e^{\beta}$	$\hat{\lambda}_0$	$\hat{\lambda}_1 = \hat{\lambda}_0 e^{\hat{\beta}}$	$\hat{\beta}$	λ_0	λ_1	β	$e^{\hat{\beta}}$
2	0.01	1.6	0.05	0.0089	0.045	1.61	-12.2	-10.0	0.7	1.1
2	0.01	3.7	0.40	0.0087	0.36	3.71	-14.7	-11.8	0.3	1.1
10	0.01	1.6	0.05	0.0099	0.050	1.61	-0.5	0.9	0.9	1.4
10	0.01	3.7	0.40	0.0099	0.45	3.82	-0.8	11.3	3.1	12.2

what would be expected based on measurement error alone. However, strong bias arises when the probability of observing a transmission in each and every couple is non-negligible.

4.6 Covariate parameter bias

Covariate parameters, such as those associated with treatment interventions, often are the primary quantities of interest in research studies, and thus it is critical to understand whether they can be accurately estimated. We use a simulation study to show that covariate parameter bias can arise in the complementary log-log transmission model in at least two important scenarios. We first consider a situation where two subgroups have infectivity rates for which a different amount of bias is expected, and second, we look at the case where two subgroups have substantially different rates of sexual activity.

For the first scenario we assume a single binary covariate x in model (4.2), and let λ_0 be the infectivity rate for the reference group ($x = 0$) and λ_1 be the infectivity rate in the second group ($x = 1$), defined as $\lambda_1 = \lambda_0 e^{\beta}$. Assuming a 50% prevalence of the covariate $x = 1$ (for example, a treatment intervention in a clinical trial) we generate 200 datasets of size 2,000 for varying levels of λ_0, ψ and β. Poisson measurement error is then applied as in (4.4), and we estimate λ_0 and β. Results are shown in Table 4.3.

Referring back to Table 4.1, when $\psi = 2$ there is almost no difference in expected bias for $\hat{\lambda}$ based on the true value of λ, and so we would not expect any bias on a covariate parameter comparing groups with different levels of infectivity. As seen in the simulation results, when $\psi = 2$ the estimated values $\hat{\lambda}_0$ are low, as expected, but the estimated values for $\hat{\beta}$ are close to the true β's regardless of the covariate effect size. (Note: $\beta = 1.6$ corresponds to a 5-fold increase in infectivity, and $\beta = 3.7$ corresponds to a 40-fold increase in infectivity.) However, when $\psi = 10$ and the effect size is large ($\beta = 3.7$), $\hat{\beta}$ is too high, with $\hat{\beta} = 3.82$. This corresponds to a 12.2% inflation of the increased risk associated with the covariate (e^{β}). These results correspond closely to Table 4.1 in that, for $\psi = 10$, no bias is observed when $\lambda = 0.01$, but 11% positive bias is expected when $\lambda = 0.40$. Using the fitted values for $\hat{\lambda}_0$ and $\hat{\beta}$ we compute $\hat{\lambda}_1 = 0.45$, which is 11.3% higher than it should be.

In the second scenario we consider a situation where two subgroups have the same level of infectivity λ (implying $\beta = 0$) but report substantially different amounts of sexual activity. Simulation results are based on datasets of size 5 000 couples with constant infectivity $\lambda = 0.005$ for all subjects. The reference group ($x = 0$) has an average of $\psi = 1$ sex act per month, while a selected subgroup ($x = 1$) has an average of $\psi = 10$ acts per month. This subgroup might be a behaviorally high-risk population, for example. Table 4.4 shows results for populations with varying proportions of the high-risk subgroup, with each row representing average results for 200 simulated datasets. Regardless of the proportion of high-risk couples in the sample we see that λ_0 is underestimated by about 26%, as expected, and consequently β is substantially overestimated at about 0.3. This would suggest that the high-risk couples have a 35% higher risk than the reference group ($e^{0.3} = 1.35$), when in fact no difference exists ($\beta = 0, e^{\beta} = 1$).

Table 4.4: Simulation results for a covariate model showing inflated covariate parameter estimates when sex frequency is not constant across subgroups and measurement error is ignored. Each estimate is the mean over 200 simulated datasets.

$\psi(x=0)$	$\psi(x=1)$	$P(x=1)$	λ_0	$\hat{\lambda}_0$	β	$\hat{\beta}$
1	10	0.1	0.005	0.0037	0	0.263
1	10	0.25	0.005	0.0038	0	0.306
1	10	0.5	0.005	0.0036	0	0.354

4.7 Controlling bias by study design

Disease infectivity is a function of the infectious agent and cannot generally be controlled by the investigator. However, the mean number of sex acts per interval could conceivably be changed by adjusting the time between encounters where individuals are queried about sexual activity and tested for disease transmission. For example, in a study of HIV prevention among heterosexual partners, we typically expect the infectivity rate to be on the order of 0.001. If the average number of sexual contacts between partners in a certain population is about one per week, then asking participants to report sex acts at one-week or one-month intervals would yield an average number of reported acts on the order of one or four, respectively. When the mean number of acts is this low we expect the simple infectivity model to yield biased estimates of λ, as seen in Tables 4.1 and 4.2. If the reporting/testing interval were expanded to eight weeks, the average number of reported acts would be high enough to prevent bias even if the mis-measurement of acts were ignored. Of course, extending recall intervals beyond a certain point can lead to increased reporting error, and so this method can only be taken so far.

Appropriate study design will depend on the disease under study. HIV-1 infection is perhaps the most extensively studied STI with respect to per-act infectivity. Estimates for female-to-male transmission range from as low as 0.001 among stable heterosexual couples (Hughes et al., 2012), to as high as 0.013 among uncircumcised men with multiple concur-

rent partners (Baeten et al., 2005). Several important risk factors appear to be positively associated with higher infectivity, including high viral loads, genital ulcerations, lack of condom use, receptive anal intercourse, and younger age (Gray et al., 2001; Wawer et al., 2005; Hughes et al., 2012; Baeten et al., 2005). Regardless of the risk group, HIV infectivity rates fall in the region of Table (4.1) where bias is only a concern if the mean number of acts ψ is small.

Other common STIs are thought to have higher infectivity rates, and therefore require more care in the selection of partner-study recall periods. For example Gray et al. (2010) found the per-act infectivity for Syphilis during insertive anal intercourse to be 0.14. Holmes et al. (1970) estimated female-to-male per-act infectivity for Gonorrhoea at 0.22 during vaginal intercourse, while Platt et al. (1983) estimated the male-to-female rate to be 0.47. In each of these cases there is very narrow range of ψ in Table 4.1 where we would expect small biases in $\hat{\lambda}$, and hence the recall window would need to be chosen with care.

When intervention effects or subgroup effects are of interest, the situation is more complicated. No longer is it enough to adjust the recall period for the sake of adjusting overall sexual frequency, but rather the infectivity rates and sexual frequency of each relevant subgroup must be considered. If the effects of a covariate (e.g. an intervention) are of overwhelming concern, whereas the actual infectivity rates are less important, the best strategy may be to choose a narrow reporting window so that ψ is small and bias is constant across multiple infectivity levels. This would minimize the risk of covariate bias. On the other hand, if investigators suspect heterogeneity in the amount of sexual activity among important subgroups, increasing ψ to a point where little bias is expected for λ in all relevant subgroups would again be critical (but not always achievable).

4.8 Discussion

Our results suggest that substantial estimation bias in both baseline infectivity and covariate effects can occur if measurement error is not accounted for in the basic infectivity model. Because the amount and direction of bias appears to depend directly on the underlying infectivity rate and on the mean number of reported sexual contacts, carefully chosen recall periods can reduce or eliminate the risk of biased estimates. In most scenarios it is enough to choose a recall period long enough to achieve a mean number of acts larger than 6. However for highly infective diseases the recall period must also be chosen so that the mean number of reported sex acts is not too large. Also, when covariate effects are of interest, the infectivity rates and sexual frequency of each subgroup must be considered.

We have also found that small-sample bias can arise when the transmission probability is high. If the chances of transmission occurring within all study couples is large enough, infectivity estimates can be dramatically overestimated. In practice, however, if either the transmission risk is so high or the observation window is so long that each and every infected individual might transmit the disease to their partner during a single observation period, a discordant partner study is not likely to be chosen as a tool for investigation. This suggests that the results presented in Table 1 for values of λ larger than 0.1, may be less relevant for STI research. Regardless, the basic transmission model is completely general, and there may exist other applications of this model where the occurrence of transmission between every pair of subjects is of practical concern.

The results presented here apply to any scenario where the regression offset term in a complementary log-log model is measured with error. For example, some non-sexually transmitted infectious diseases may be thought of as having constant infection risk, with the probability of infection modeled as

$$P(y = 1) = 1 - e^{-\beta t} \tag{4.7}$$

where t is the exposure time and β is the infection rate per unit time (Draper et al., 1972). Applying the complementary log-log transformation to this model yields a convenient linear form, where $log(\beta)$ is the regression intercept with $log(t)$ as the offset. If exposure time is measured with error, $\hat{\beta}$ may suffer estimation bias similar to that of per-act infectivity λ as reported here.

An important assumption that we make in our simulation procedure is that the misreporting of sex acts is itself unbiased. If, in fact, the reported number of acts is higher on average than the true number of acts, then our estimate of infectivity would tend to be too low. The opposite would be true for under reporting of sexual activity. In general, we do not have reason to suspect an over or under reporting of sexual activity, although Clark and Wallin (1964) provide compelling evidence that biased reporting can happen in certain subgroups. For example, we might expect underreporting when subjects are queried about sex acts following abstinence counseling, but we don't attempt to model those issues here.

A better understanding of sexual-history misreporting would certainly inform our understanding of potential estimation bias, but validity and reliability data are not routinely collected in HIV-affected populations. Recently, however, new studies have begun collecting sexual history data from both partners. Discrepancies between reports from the partners could shed light on the nature of mis-reporting distributions.

In our analysis, the assumed distributions of true sex acts and measurement error are only rough approximations. Our recommendations for discordant-couple studies assume that sex acts and measurement error follow a roughly Poisson or Negative Binomial shape, however if the true distributions were substantially different, the expected estimation bias

would likely be different as well. Provided that our assumptions are reasonably close to reality, a rule of thumb suggested by our results is that for studies of low-infectivity STI's, recall intervals should be long enough that the expected mean number of acts is not small. This interval could vary dramatically depending on the population under study, and should not be chosen so long that subjects cannot reasonably be expected to remember the number of sexual encounters. In addition, estimated covariate parameters comparing subgroups with substantially different infectivity rates or substantially different sexual frequencies should be interpreted with caution.

In situations where it is not possible to modify the study design so as to avoid estimation bias, statistical modeling may offer helpful methods. Although the use of measurement-error models is now commonplace, we are not aware of any techniques that apply to this particular setting. There is a need for partner-study designs and analytic methods that account for misreported sexual activity in order to prevent estimation bias. In future work we will propose such methods.

4.9 Appendix A

The expected value of the derivative of the log likelihood (4.5) is given by

$$\begin{aligned}
\mathrm{E}[\frac{d}{d\lambda}\log(f(\boldsymbol{y}|\lambda,\boldsymbol{m}))] &= \sum_{i=1}^{N}\mathrm{E}\left[\frac{d}{d\lambda}\log[\{1-(1-\lambda)^{m_i}\}^{y_i}\{(1-\lambda)^{m_i}\}^{1-y_i}]\right] && (4.8)\\
&= N\times\mathrm{E}\left[\frac{d}{d\lambda}\log[\{1-(1-\lambda)^{m}\}^{y}\{(1-\lambda)^{m}\}^{1-y}]\right] && (4.9)\\
&= N\sum_{y=0}^{1}\sum_{m=1}^{\infty}\left[\frac{ym(1-\lambda)^{m-1}}{1-(1-\lambda)^{m}}-\frac{m(1-y)}{1-\lambda}\right]f(y,m|\lambda) && (4.10)
\end{aligned}$$

where $f(y,m|\lambda)$ is the joint density of y and m, marginalized over the true number of acts n. When $n\sim$ Poission(ψ) and $m\sim$ Poisson(n^*) we have

$$f(y, m|\lambda) = \sum_{n=0}^{\infty} f(y|n, \lambda) f(m|n^*) f(n|\psi) \tag{4.11}$$

$$= \sum_{n=0}^{\infty} (1 - (1-\lambda)^n)^y (1-\lambda)^{n(1-y)} \frac{n^{*m} e^{-n^*}}{m!} \frac{\psi^n e^{-\psi}}{n!} \tag{4.12}$$

where $n^* = 0.25$ if $n = 0$ and $n^* = n$ otherwise.

Chapter 5

ESTIMATING PER COITAL ACT INFECTIVITY WHEN SEXUAL ACTIVITY IS REPORTED WITH ERROR.

5.1 Abstract

Discordant partner studies are commonly used to quantify per-act infectivity rates for HIV. However these studies depend on self-reported sexual histories which are notoriously unreliable and can yield biased parameter estimates. With the hope of improving estimation, recent discordant partner studies have begun collecting sexual history data from both partners. We propose a transmission model which incorporates data from both partners, and accounts for the mis-reporting of sex acts as well as non-overlapping recall periods. We use simulation studies to demonstrate that this new method can dramatically reduce the measurement-error bias for per-act infectivity estimation. We then apply this method to the Partners PrEP study dataset and show that HIV infectivity may be 25% higher than would be estimated with traditional models.

5.2 Introduction

Each sexual contact between an individual infected with a sexually transmitted infection (STI) and an uninfected individual may result in disease transmission. The likelihood of such a transmission is referred to alternately as the "per-act infectivity" or "per-act transmission probability", and this quantity is of central interest in the study of STI prevention and epidemic modeling. The infectivity rate is needed for modeling infection dynamics, identifying infection risk factors, and designing prevention strategies. Also, differences in per-act infectivity can serve as a primary measure of success or failure of STI-prevention interventions.

A number of published studies provide estimates of per-act infectivity for the most common types of sexual activity. A recent meta-analysis (Patel et al., 2014) of studies in high-income countries estimates the per-act risk to be 0.0004 for insertive vaginal intercourse (male-to-female) and 0.0008 for receptive vaginal intercourse (female-to-male). The risk associated with anal intercourse is estimated to be higher, at 0.0011 for insertive anal intercourse and 0.014 for receptive anal intercourse. It is important to note that these aggregate estimates are based on populations that range in their composition with respect to important predictors of infectivity, such as viral load, condom use, male circumcision, concurrent infections, and anti-retroviral therapy (ART) use. A somewhat older meta-analysis (Boily et al., 2009) of studies in low-income settings found the heterosexual rates to be higher, at 0.0030 for male-to-female and 0.0038 for female-to-male. However, the size and quality of the these studies range broadly.

Prospective discordant-couple studies are a way to study the risk of disease transmission in a relatively controlled environment where the HIV status of both partners in known and sexual activity is monitored regularly. In a typical discordant-partner study, one member of a partnership is interviewed at regular intervals regarding sexual activity. Simultaneous to the interview, the disease-negative partner is tested to determine whether transmission has occurred. By counting the number of sexual contacts between each pair of partners prior to transmission, it is possible to estimate the per-act infectivity using the simple transmission model proposed by Jewell and Shibowski (1990). Hughes et al. (2012) extended this model to account for condom use and other important covariates.

All models of per-act infectivity rely on self-reported sexual activity. It has been shown, however, that retrospective reports of the number of sexual encounters are often inaccurate (Clark and Wallin, 1964). It also has been shown that the longer the time span over which people are asked to quantify their sexual activity, the less reliable those reports become

(Kauth et al., 1991). Thus, even in the relatively controlled setting of prospective longitudinal discordant-partner studies, accurately quantifying sexual activity is a challenge.

It is well known that covariate measurement error and misclassification can lead to severely biased regression parameter estimates in linear and non-linear models (Barron, 1977; Fuller, 1987; Carroll et al., 2006). In the context of an STI transmission model, Hanscom et al. (2014b) showed that sex-act misreporting, when ignored, can lead to substantial estimation bias (positive and negative) for both per-act infectivity and covariate parameter effects. Various methods have been proposed to correct for the effects of mismeasured variables, for example obtaining accurate, gold-standard measures of exposure on a validation sample (Hotchberg, 1977; Morrissey and Spiegelman, 1999; White et al., 2001; Carroll et al., 2006). There is no gold standard for collecting accurate sexual histories, however, and therefore gold-standard type methods (e.g regression calibration) are not appropriate.

Another strategy involves obtaining repeated measurements of error-prone variables for some or all individuals. Recently, large discordant partner studies have begun collecting sexual history data from both partners, and these replicate assessments lend themselves to dual-reporting methods. A number of methods have been proposed that utilize repeat assessments for measurement-error correction (Rosner et al., 1992). Dalen et al. (2009) describe a non-parametric model that estimates the true exposure distribution conditional on replicate observations and outcomes, however this method does not accommodate covariates and only applies to estimated odds ratios. Devanayaran and Stefanski (2002) extend the simulation extrapolation approach (SIMEX) to replicate report data, however SIMEX methods (Cook and Stefanski, 1995; Wang et al., 1999) assume the measurement error variance is known, and it is not clear that we can find an appropriate extrapolation function for SIMEX in the context of the extended transmission model we use in this paper.

Latent-class models have been proposed for use with replicate binary exposure data in epidemiologic association studies (Kaldor and Clayton, 1985; Evans et al., 1989) and in studies of the value of diagnostic markers (Rindskopf and Rindskopf, 1986). Provided that replicate measures are independent conditional on the true exposure, unbiased estimates of exposure effects can be obtained in a variety of settings. Similarly, latent-variable models can be used for continuous measurement error situation (Thomas, 2007), and both latent-class and latent-variable models are often fit using expectation maximization (EM) (Gustafson, 2004). For complex models, however, the expectation step (E-step) in the EM procedure can require challenging numerical approximations (Schafer, 2002), and a convenient, flexible alternative is to fit these models using a Bayes-MCMC approach (Richardson and Gilks, 1993a,b; Gustafson, 2004; Carroll et al., 2006).

One important complication that arises in dual-reporting HIV partner studies is that the partners are not always interviewed on the same day, and hence their reports of sexual activity do not always correspond to the same time periods. Existing replicate-measurement methods have not been designed to handle this partial-replication situation. In this paper we propose a model that incorporates a novel latent-variable structure which explicitly accounts for non-overlapping recall intervals. Our proposed model also incorporates a flexible measurement-error distribution which can handle varying degrees of sex-act misclassification, and we use Bayes-MCMC for fitting and inference.

5.3 Modeling per act infectivity

5.3.1 The basic transmission model and extensions

A basic infectivity model was described by Jewell and Shibowski (1990) in the context of HIV partner studies. This model is derived from a typical time-to-event framework, but

uses the number of sexual acts n in place of a measure of time, thus allowing for estimation of the per-act infectivity rate λ. Letting y be 1 if a transmission occurred after n sexual contacts, and 0 otherwise, the probability of transmission can be expressed as

$$P(y = 1) = 1 - (1 - \lambda)^n. \tag{5.1}$$

The model assumes constant infectivity over time, and can be expanded to accommodate covariates X as follows,

$$P(y = 1) = 1 - (1 - \lambda)^{ne^{X\beta}} \tag{5.2}$$

When λ is small equation (5.2) is approximately equal to

$$1 - (1 - \lambda e^{X\beta})^n \tag{5.3}$$

and therefore β is approximately the log relative risk of per-act transmission per unit change in X. Estimation can be achieved using standard GLM optimization under the complementary log-log link function (Jewell and Shibowski, 1990). The model has been further extended by Hughes et al. (2012) to properly account for condom use which, because it is a variable that changes within the reporting period, must be handled differently than covariates that are assessed only once per reporting period. This model defines two infectivity parameters, λ_1 for infectivity when condoms are used and λ_0 for infectivity when condoms are not used. The model is as follows:

$$P(y = 1) = 1 - (1 - \lambda_0)^{n_0 e^{X\beta}} (1 - \lambda_1)^{n_1 e^{X\beta}}. \tag{5.4}$$

where n_0 is the number of acts where condoms were not used and n_1 is the number of acts where a condom was used. Covariates parameters β are assumed to be the same for both types of acts, however this assumption could be relaxed. The added terms in this model preclude the use of a complementary log-log link function to create linearity in the covariate parameters, and hence standard GLM estimation is not possible. When we introduce the latent variables and the measurement error distribution the model becomes even more complex, and so we turn to a Bayesian framework and use MCMC for parameter estimation.

An added feature of model (5.4) is that it permits an estimate of how often condom use is over-reported. If we assume that the probability of HIV transmission is effectively zero when a condoms are properly used (i.e. $\lambda_1 = 0$), and we also assume that some proportion c of acts reported with condoms (n_1) did not actually involve a condom (or perhaps the condom broke or was used improperly), we can reformulate model (5.4) as:

$$P(y = 1) = 1 - (1 - \lambda_0)^{(n_0 + cn_1)e^{X\beta}}. \tag{5.5}$$

Models (5.4) and (5.5) have the same number of parameters and are algebraically equivalent when $c = log(1-\lambda_1)/log(1-\lambda_0)$. As a result we have a convenient estimator of the fraction of acts where condom-use is over-reported, namely $\hat{c} = log(1 - \hat{\lambda}_1)/log(1 - \hat{\lambda}_0)$, where $\hat{\lambda}_1$ and $\hat{\lambda}_0$ are estimated from (5.4). We assume that the cn_1 acts are equivalent to acts with no condom use in terms of HIV transmission risk. The quantity c can be thought of as the probability that an act reported with condom use did not in fact involve a properly used, intact condom.

5.3.2 A latent-variable transmission model with measurement-error and non-overlapping time intervals

Consider N serodiscordant couples monitored monthly over a maximum of two years. Let $\boldsymbol{T1}_i$ be a vector of times, measured in days from enrollment for couple i, when the HIV- partner was interviewed and tested, and let $\boldsymbol{T2}_i$ be a vector of times indicating when the HIV+ index was interviewed. At each interview the subject is asked to report the number of sex acts that occurred during the previous w days, and further indicate during how many of those acts a condom was used. Let $\boldsymbol{t}_i$ be the ordered, combined set of all visit times $\boldsymbol{T1}_i$ and $\boldsymbol{T2}_i$ for couple i, as well as the set of all time points w days prior to the interview dates, with duplicates removed. Each element t_{ij} of $\boldsymbol{t}_i$ thus represents either the beginning or end of a reporting window, and the time intervals defined by $\boldsymbol{t}_i$ represent a maximal set of time periods relevant to all available reported sexual-act data for couple i. (See Figure 1.)

Let $\boldsymbol{m1} = \{m1_{ivl}\}$ be a matrix of the reported number of sex acts where a condom was used, as reported by partner l of couple i at visit v. Similarly let $\boldsymbol{m0} = \{m0_{ivl}\}$ be the matrix of reported sex acts where a condom was not used, as reported by partner l of couple i at visit v. All reported counts are assumed to be independent conditional on the true numbers of acts. Each reported count is associated with one or more time intervals as defined by $\boldsymbol{t}$. Let $\boldsymbol{y}_i$ be the vector of binary (0/1) HIV test results for the HIV– partner in couple i. Test results only exist at time points where the HIV– partner was interviewed and tested.

Let $\boldsymbol{n}_i$ be the vector of true (unobserved) act counts for couple i during time intervals $\boldsymbol{t}_i$, and let $\boldsymbol{n1}_i$ be the number of those acts where a condom was used. Both $\boldsymbol{n}$ and $\boldsymbol{n1}$ are latent variables; while they measure the true numbers of acts, we cannot observe them directly. Note that each reported act count $m0_{ivl}$ and $m1_{ivl}$ refer to a w-day time window, and will be associated with one or more true act counts n_{ij} which are defined on the maximal

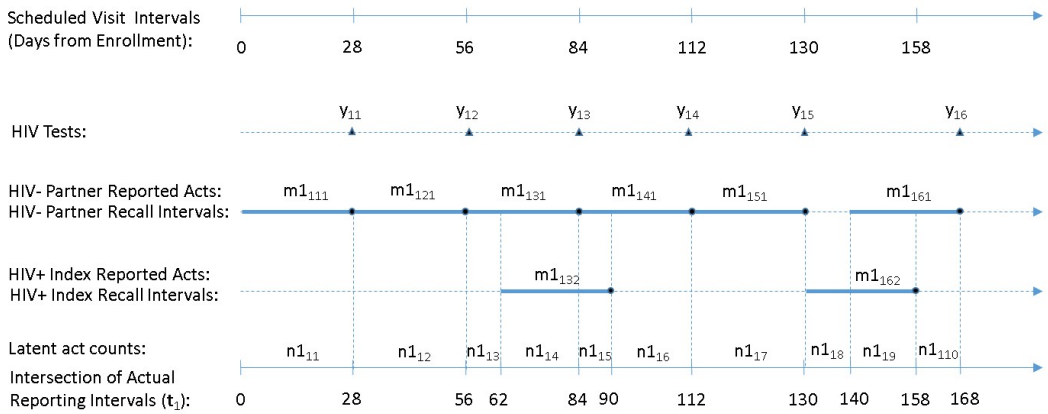

Figure 5.1: Schematic representation of scheduled and actual visit times for an example discordant couple. The thick line segments represent the four week (28 day) recall window for partner and index act reporting. In this example, index cases are only interviewed once every three months, and for simplicity only acts with reported condom use are shown. At the first two scheduled visits the HIV– partner is interviewed exactly on time, and their reported acts with condom use ($m1_{111}$, $m1_{121}$) correspond exactly with the latent counts ($n1_{11}$, $n1_{12}$). The reported quantity $m1_{111}$ is a potentially incorrect version of the true number of acts $n1_{11}$, and $m1_{121}$ is the reported version of the true quantity $n1_{12}$. The probability of HIV transmission during those intervals ($\mathrm{P}[y_{11} = 1]$, $\mathrm{P}[y_{12} = 1]$) is associated with the true act counts $n1_{11}$ and $n1_{12}$, respectively. At the third visit the index case is interviewed six days late, and the intersection this reporting window with the reporting windows for the HIV- partner yield three sub-intervals and the corresponding latent counts $n1_{13}$, $n1_{14}$, and $n1_{15}$. The act count reported by the HIV– partner at visit three ($m1_{131}$) is treated as an error-prone report of the true act count $n1_{13} + n1_{14}$. Similarly, the number acts reported by the Index case at time three ($m1_{132}$) is an error-prone report of the true quantity $n1_{15} + n1_{15}$. The probability of transmission during the third interval ($\mathrm{P}[y_{13} = 1]$) is associated with the true number of acts $n1_{13} + n1_{14}$. At month six the HIV– Partner is interviewed 10 days behind schedule, again inducing three sub intervals and corresponding latent quantities $n1_{18}$, $n1_{19}$, and $n1_{110}$. In this case the probability of HIV transmission at test six ($\mathrm{P}[y_{16} = 1]$) is dependent upon the true number of acts $n1_{18} + n1_{19} + n1_{110}$.

set of time intervals. Figure 5.1 is a schematic representation of observed visit times and reporting windows for an example discordant couple. The intersection of actual reporting windows are displayed along with the corresponding latent counts. Variables and variable indexing are defined in Table 5.1.

Equations (5.6) through (5.15) define the full data model. The transmission probability model (5.7) is equivalent to (5.4) for λ_0 and λ_1 small. Our measurement-error distribution for reported acts is specified as a mixture of three components, and accounts for not only mis-reporting, but also the possibility of erroneous outliers that are uncorrelated with reality (Equations (5.12) and (5.13)). The first component represents the case where the reported number of acts is exactly accurate. The second component represents the case where acts are reported with error, and the error distribution is Poisson with a median equal to the true number of acts. The third component represents outliers which have little or no association with the true number of acts. Probabilities for each of these three cases are represented by the vector $\boldsymbol{p} = \{p_1, p_2, p_3\}$. We are assuming non-differential measurement error, which implies that conditional on the true act counts, the reported act counts are independent and uncorrelated with y. Also note that because we are specifying a distribution for the latent variables in (5.10) we are using a 'structural' modeling approach, as defined by Carroll et al. (2006).

variable	values	definition
Observed Data		
y_{ik}	binary 0/1	HIV test results (1=positive) for HIV- partner, couple i, test k
t_{ij}	days (positive)	time from enrollment, couple i, interval j, where $t_{i0} = 0$
$m1_{ivl}$	count	reported number of acts with condom, couple i, visit v, partner l
$m0_{ivl}$	count	reported number of acts without condom, couple i, visit v, partner l
$\boldsymbol{x}_{ik}$	any	vector of covariate values, either partner, couple i, test visit k
Latent Variables		
n_{ij}	count	true/total number of acts, couple i, interval j
$n1_{ij}$	count	true number acts with condom, couple i, interval j
Unknown Parameters		
λ_1	probability	per-act infectivity with condom
λ_0	probability	per-act infectivity w/o condom
$\boldsymbol{\beta}$	real	vector of covariate parameters (log RR)
$\boldsymbol{p}$	probability	vector of measurement-error mixing probabilities
Random effects		
ψ_i	rate	average w-day sex act rate, couple i
$p.condom_i$	probability	probability of using a condom, couple i
Index Variables		
$VisitInt_{ivl1}$	integer	first interval associated with vth visit, couple i, partner l
$VisitInt_{ivl2}$	integer	last interval associated with vth visit, couple i, partner l
$TestInt_{ik1}$	integer	first interval associated with kth test, couple i
$TestInt_{ik2}$	integer	last interval associated with kth test, couple i

index	indexing	range	maximum
i	couple	$1...N$	N = number of couples
j	interval	$1...J_i$	J_i = number of intervals for couple i
v	visit	$1...V_{il}$	V_{il} = number of visits for couple i, partner l
l	partner	{1,2}	1=HIV+ index case, 2=HIV– partner
k	HIV test visit	$1...TV_i$ for couple i	TV_i = number of visits with HIV test data

Table 5.1: Variables and indices for latent variable transmission model.

$$
\begin{aligned}
y_{ik} &\sim \text{Bernoulli}(r_{ik}) && (5.6)\\
r_{ik} &= 1-(1-\lambda_0\exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1\exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}} && (5.7)\\
n1'_{ik} &= \sum_{j=TestInt_{ik1}}^{TestInt_{ik2}} n1_{ij} && (5.8)\\
n0'_{ik} &= \sum_{j=TestInt_{ik1}}^{TestInt_{ik2}} n_{ij}-n1_{ij} && (5.9)\\
n_{ij} &\sim \text{Poisson}(\psi_i(t_{ij}-t_{i(j-1)})/w) && (5.10)\\
n1_{ij} &\sim \text{Binomial}(n_{ij}, p.condom_i) && (5.11)\\
m1_{ivl} &\sim p_1 1_{[m1_{ivl}=n1_{ivl}]}+p_2\text{Poisson}(n1^*_{ivl})+\frac{1}{100}p_3 1_{[m1_{ivl}\geq 29]} && (5.12)\\
m0_{ivl} &\sim p_1 1_{[m0_{ivl}=n0_{ivl}]}+p_2\text{Poisson}(n0^*_{ivl})+\frac{1}{100}p_3 1_{[m0_{ivl}\geq 29]} && (5.13)\\
n1^*_{ivl} &= \max(0.25, \sum_{j=VisitInt_{ivl1}}^{VisitInt_{ivl2}} n1_{ij}) && (5.14)\\
n0^*_{ivl} &= \max(0.25, \sum_{j=VisitInt_{ivl1}}^{VisitInt_{ivl2}} n_{ij}-n1_{ij}) && (5.15)
\end{aligned}
$$

5.4 Bayesian Estimation of per-act Infectivity and Covariate Parameters

5.4.1 The Likelihood

The joint likelihood for all parameters can be expressed as

$$
\begin{aligned}
&L(\lambda_0,\lambda_1,\boldsymbol{\beta},\boldsymbol{\psi},\boldsymbol{p.condom},\alpha_1,\alpha_2,\nu_1,\nu_2;\boldsymbol{y},\boldsymbol{m1},\boldsymbol{m0},\boldsymbol{x},\boldsymbol{t}) =\\
&\quad \sum_{n}\sum_{n1} p(\boldsymbol{y}|\lambda_0,\lambda_1,\boldsymbol{n},\boldsymbol{n1},\boldsymbol{x},\boldsymbol{\beta})p(\boldsymbol{n}|\boldsymbol{\psi},\boldsymbol{t})p(\boldsymbol{n1}|\boldsymbol{n},\boldsymbol{p.condom})\\
&\times\ p(\boldsymbol{m1}|\boldsymbol{n1})p(\boldsymbol{m0}|\boldsymbol{n},\boldsymbol{n1})p(\boldsymbol{\psi}|\alpha_1,\alpha_2)p(\boldsymbol{p.condom}) \qquad (5.16)
\end{aligned}
$$

$$
\begin{aligned}
= & \prod_{i=1}^{N}\prod_{k=1}^{TV_i}\left[1-(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}} \\
\times & \left[(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}} \\
\times & \prod_{j=1}^{J_i}\frac{1}{n_{ij}!}(\psi_i(t_{ij}-t_{i(j-1)})/28)^{n_{ij}}exp(-\psi_i(t_{ij}-t_{i(j-1)})/28) \\
\times & \prod_{j=1}^{J_i}\binom{n_{ij}}{n1_{ij}}p.condom_i^{n1_{ij}}(1-p.condom_i)^{n_{ij}-n1_{ij}} \\
\times & \prod_{l=1}^{2}\prod_{v=1}^{V_{il}}p_1 1_{[m1_{ivl}=n1_{ivl}]}+p_2 n1_{ivl}^{*m1_{ivl}}exp(-n1^*_{ivl})+\frac{1}{100}p_3 1_{[0\le m1_{ivl}\le 99]} \\
\times & \prod_{l=1}^{2}\prod_{v=1}^{V_{il}}p_1 1_{[m0_{ivl}=n0_{ivl}]}+p_2 n0_{ivl}^{*m0_{ivl}}exp(-n0^*_{ivl})+\frac{1}{100}p_3 1_{[0\le m0_{ivl}\le 99]} \\
\times & \prod_{i=1}^{N}\psi_i^{\alpha_1-1}exp(-\alpha_2\psi_i)\times\prod_{i=1}^{N}p.condom_i^{\nu_1-1}(1-p.condom_i)^{\nu_2-1} \qquad (5.17)
\end{aligned}
$$

The presence of latent variables combined with random effects renders maximum likelihood parameter estimation computationally difficult (Wannemuehler, 2005). Instead we use a Bayesian approach and Markov Chain Monte Carlo (MCMC) simulation for estimation and inference. The Bayes-MCMC approach offers tremendous flexibility and is relative easy to implement even for relatively complex models (Carroll et al., 2006).

5.4.2 Prior Distributions - General Form

The general forms of the prior distributions are as follows. Specific distributions can be chosen depending on the particular dataset being modeled. Beta (or Uniform) distributions are used for all variables whose range is zero to one, Gamma distributions for variables whose range is the positive real line, and Normal distributions for the covariate parameters whose possible values are positive or negative. Each of these choices may be made more or less informative depending on the quality of prior information. Similarly, the Dirichlet distribution is a natural choice for mixing probabilities, and prior variances may be chosen

as appropriate. Hyperparameters $(\nu_1, \nu_2, \alpha_1, \alpha_2)$ can be pre-specified or estimated from the data.

$$\begin{aligned}
\lambda_0 &\sim \text{Beta}(\phi_1, \phi_2) && (5.18)\\
\lambda_1 &\sim \text{Beta}(\tau_1, \tau_2) && (5.19)\\
\boldsymbol{\beta} &\sim \text{MVN}(0, \Sigma = \text{diag}(\boldsymbol{\sigma}^2)) && (5.20)\\
p.condom_i &\sim \text{Beta}(\nu_1, \nu_2) && (5.21)\\
\nu_1 &\sim \text{Beta}(\mu_1, \mu_1) && (5.22)\\
\nu_2 &\sim \text{Beta}(\kappa_1, \kappa_2) && (5.23)\\
\psi_i &\sim \text{Gamma}(\alpha_1, \alpha_2) && (5.24)\\
\alpha_1 &\sim \text{Gamma}(\delta_1, \delta_2) && (5.25)\\
\alpha_2 &\sim \text{Gamma}(\omega_1, \omega_2) && (5.26)\\
\boldsymbol{p} = \{p_1, p_2, p_3\} &\sim \text{Dirichlet}(\gamma_1, \gamma_2, \gamma_3) && (5.27)
\end{aligned}$$

5.4.3 Parameter Estimation

Parameter estimation is achieved using MCMC sampling from the posterior distribution of the parameters. The set of full conditional distributions needed for the sampling are listed in Appendix A, along with an outline of the Gibbs sampling procedure.

5.5 Simulation Example

We conducted simulation studies to test how well the model performs in realistic scenarios with missing data and with incorrectly specified distributions for sex-act mispreporting. Example data were simulated so as to match the observed data in the Partners PrEP dataset (Section 5.6) as closely as possible and hence provide a close comparison to a real application.

5.5.1 Simulated Datesets

The reporting interval was defined as four weeks, and the couple-specific act rate ψ_i for these intervals was generated as Gamma(2,0.4), giving a mean act rate of 5.0. Couple-specific reported condom use rates $p.condom_i$ were generated as Beta(0.5, 0.01), yielding a reported condom use-rate of about 85%. Actual interval lengths were randomly offset from 28 days to match the reality that many people are not interviewed on their intended visit date, and partners within the same couple are not always interviewed on the same day. Reported act counts were generated according to models (5.12) and (5.13) with $\boldsymbol{p}$ = (0.4975, 0.4955, 0.005). Transmission probability was associated with two covariates: treatment and viral load. Fifty percent of couples were placed in the treatment group, and the log relative risk of transmission (β_1) was -0.69, corresponding to a relative risk of 0.50. Viral load was generated as Gamma(0.7, 0.00002) and was centered and included in the model on the log10 scale with log relative risk β_2=1.1. This corresponds a relative risk of 3.0 per additional log10 viral RNA copies per mL. Per-act infectivity was assumed to be 0.0002 for acts with reported condom use and 0.004 without reported condom use.

5.5.2 Prior Distributions

In order that key parameter estimates are based as much on observed data as possible, minimally informative priors were chosen for the key parameters of interest:

$$\lambda_0 \sim \text{Unif}(0, 0.01) \tag{5.28}$$

$$\lambda_1 \sim \text{Unif}(0, 0.01) \tag{5.29}$$

$$\boldsymbol{\beta} \sim \text{MVN}(0, \text{diag}(10)). \tag{5.30}$$

Couple-specific act rates ψ_i are distributed as Gamma(α_1, α_2) with priors $\alpha_1 \sim$ Gamma(4,2)

and $\alpha_2 \sim$ Gamma(4,10). This gives an induced prior mean for ψ_i corresponding to the true mean of 5.0, but allows for considerable flexibility on the posterior estimates. The prior for $\boldsymbol{p}$ is Dirichlet(0.495, 0.495, 0.01) which is centered on the true values.

5.5.3 Parameter Estimation

The MCMC procedure outlined in Appendix A was carried out using *JAGS* software (Version 3.4.0) via the *R* interface for 100 simulated datasets. 15,000 MCMC iterations were sampled, and the first 10,000 samples were dropped as burn in. Each of the 100 simulations took approximately 24 hours to run. (Appendix B contains details regarding convergence diagnostics.)

5.5.4 Comparison Model and Results

To evaluate the effectiveness of our proposed model, we compare it to a model that does not account for act mis-reporting or mis-aligned reporting windows. The full model, as described in section (5.3.2) and a 'naive' model were both fit to each of the 100 simulated simulated datasets. The naive model assumes that the reported numbers of acts, as reported by the HIV-negative partner, are in fact the true number of acts, and ignores HIV-positive partner report. The same prior distributions are used as in the full model, with the exception $\boldsymbol{p}$ which does not appear in the naive model since no mis-reporting is assumed. Both of these models are compared to the results of a model that is fit using the true, underlying numbers of acts, again using the same priors.

Mean estimates over the 100 simulations are included in Table 5.2, as well as percent bias. The naive model underestimates λ_0 by 40%, while the full model is off by only 0.3%. Similarly the naive model underestimates λ_1 by 5.5%, while the full model is off by 2.7%. Both models perform similarly with respect to the covariate parameter estimates, missing

the mark by about one percent or less.

	Simulation Values	Estimates using true act counts	Full model Estimates	Percent Bias	Naive model Estimates	Percent Bias
λ_0	0.002	0.00219	0.00220	0.3	0.00131	-40.2
λ_1	0.0004	0.000408	0.000397	-2.7	0.000386	-5.5
β_1	-0.69	-0.735	-0.737	-0.2	-0.728	1.0
β_2	1.1	1.106	1.104	-0.1	1.091	-1.3
α_1	2.0	1.99	1.99	-0.4	2.12	6.1
α_2	0.4	0.399	0.388	-2.7	0.361	-9.4
p_1	0.4975	-	0.491	-	-	-
p_2	0.4975	-	0.503	-	-	-
p_3	0.0005	-	0.0059	-	-	-

Table 5.2: Mean estimates for 100 simulated datasets, and percent bias as compared to estimates based on true act counts. Parameter estimates for individual simulations are posterior sample means from the MCMC sequence, excluding burn in.

5.5.5 Overdispersed Measurement Error

To further test robustness of model mis-specification, a second simulation study was conducted where the true measurement-error distribution was not the same as the distribution assumed by the model. We imagine a situation where the reported number of acts deviates from the true number of acts by a larger amount than would be predicted by the Poisson model (i.e. overdispersion). The triple mixture was still used as in equations (5.12) and (5.13), but in place of the Poisson we use a negative binomial distribution with the same mean but broader variance:

$$m1_{ivl} \sim p_1 1_{[m1_{ivl}=n1_{ivl}]} + p_2 \text{NegBinom}(\text{mean} = n1^*_{ivl}, \text{size} = n1^*_{ivl}) + \frac{1}{100} p_3 1_{[m1_{ivl} \geq 29]} \quad (5.31)$$

$$m0_{ivl} \sim p_1 1_{[m0_{ivl}=n0_{ivl}]} + p_2 \text{NegBinom}(\text{mean} = n0^*_{ivl}, \text{size} = n0^*_{ivl}) + \frac{1}{100} p_3 1_{[m0_{ivl} \geq 29]} \quad (5.32)$$

The negative binomial portion of this mixture has measurement error variance about twice that of a Poisson with the same mean. Results are displayed in Table 5.3. For the naive model, overdispersed measurement error again resulted substantial bias for both λ_0 and λ_1 (-39.3% and -6.5% respectively). The full model provides correction for this bias to within 2-3%, giving estimates for λ_0 and λ_1 very close to those that would be obtained if the true numbers of acts were known. Covariate parameter estimates are virtually unbiased for the full model and slightly biased for the naive model. These results suggest that the proposed model is robust to misspecification of the measurement error model.

	Simulation Values	Estimates using true act counts	Full model Estimates	Percent Bias	Naive model Estimates	Percent Bias
λ_0	0.002	0.00207	0.00213	2.9	0.00126	-39.3
λ_1	0.0004	0.000414	0.000406	-2.0	0.000387	-6.5
β_1	-0.69	-0.742	-0.743	-0.1	-0.734	1.1
β_2	1.1	1.079	1.077	-0.2	1.055	-2.3
α_1	2.0	1.99	1.89	-4.9	2.09	4.8
α_2	0.4	0.397	0.373	-6.2	0.355	-10.7
p_1	0.497	-	0.418	-	-	-
p_2	0.497	-	0.574	-	-	-
p_3	0.0005	-	0.0083	-	-	-

Table 5.3: Mean parameter estimates and percent bias for 80 simulated datasets/analyses. The true measurement error distribution is negative binomial, while the assumed measurement error distribution (in the full model) is Poisson. The naive model assumes no measurement error. Estimates are observed means based on MCMC iterations 10,000 through 15,000.

5.6 Parters PrEP Study Example

The Partner's in Prevention Pre-Exposure Prophylaxis (PrEP) study enrolled 4,758 HIV1 serodiscordant couples between 2008 and 2010 in Uganda and Kenya (Baeten et al., 2012). The HIV negative partner was randomly assigned to one of three treatment regimens - placebo, tenovofir (TDF), or combination tenofovir and emtricitabine (TDF-FTC), and the couples were followed for a maximum of 36 months. The HIV negative partner was queried about sexual activity every four weeks, and also tested for HIV acquisition. The HIV positive partner was queried once every three months, and also tested for HIV viral load. We analyzed the PrEP study data using both the full model and the naive model, with both models using the same priors, the same starting values, and 40,000 MCMC iterations with 30,000 dropped as burn in.

Prior distributions were selected to be minimally informative for primary parameters and somewhat informative for secondary parameters (Table 5.4.) Based on prior estimates summarized by Boily et al. (2009) and Hughes et al. (2012) we expect per-act infectivity to be between 0.0001 and 0.005, and hence we use flat priors on the range (0, 0.01). These priors allow for a reasonable range of possible values, but prevent the MCMC sampler from straying towards highly unlikely estimates. Priors for the covariate parameters are intended to be nearly flat over a broad range of positive and negative values, allowing the posterior to be determined primarily by observed data.

Reported act counts in the PrEP study suggest an average monthly sex act rate of about 5.0, and hence we let α_1 have prior mean 2.0 and α_2 have prior mean 0.4, giving a mean of 5.0 for the couple-specific mean act rate ψ_i. Observed data from the PrEP study also provide information regarding the measurement-error mixture parameters. Considering only intervals where both partners are interviewed on the same day, about 50% of the time both partners report identical act counts. We therefore set the prior mean for p_1 at about 0.5.

Act reports from these same data suggest that a small percentage (less than 1%) of act reports are wildly mistaken or data-entry errors. (For example, there are cases where one partner reports less than 5 acts during the past month and the other partner reports more than 30.) We therefore use a prior mean for p_3 of about 0.01. The remaining 50% of reports are assumed to follow approximately Poisson error, and the prior mean for p_2 is set at about 0.5. Using the Dirichlet prior assures that the mixing parameters sum to one, and using prior parameters (50,50,1) gives a moderate prior standard deviation for all elements of $\boldsymbol{p}$ (Figure 5.18, Appendix C).

Parameter	Prior
λ_0	Unif(0,0.01)
λ_1	Unif(0,0.01)
β_1 (TDF)	N(0,100)
β_2 (TDF-FTC)	N(0,100)
β_3 (HIV+ Male)	N(0,100)
ψ_i	Gamma(α_1, α_2)
α_1	Gamma(4,2)
α_2	Gamma(4,10)
$\boldsymbol{p}$	Dirichlet(50, 50, 1)

Table 5.4: Prior distributions for PrEP analysis parameters.

Convergence diagnostics are included in Appendix C. Complete traceplots shown in Figures 5.9 and 5.10 show that convergence is rapid for the primary parameters of interest (infectivity rates and covariate parameters), but much slower for the secondary parameters (hyperparameters α_1 and α_2, and the measurement-error mixing parameters). Figures 5.11 and 5.12 display the final 10,000 MCMC iterations, and appears to show a good level of convergence for all parameters except the mixing parameters $\boldsymbol{p}$ which don't appear to converge completely. Cumulative mean and autocorrelation plots show a similar picture of convergence (Figures 5.13-5.16), with primary parameters showing solid convergence and mixing parameters still not entirely converging. The posterior distributions of the model

parameters shown in Figures 5.17 and 5.18 suggest that neither the prior distributions nor the starting values appear to have unduly influenced the posterior estimates, and the sampled posterior distributions are tightly focused. Lack of convergence by the MCMC chains for $\boldsymbol{p}$ is concerning, however $\boldsymbol{p}$ is a essentially a nuisance parameter and the fact that we have good convergence for our primary parameters is reassuring.

Parameter estimates (posterior means) and credible intervals for both the full and naive models are displayed in Table 5.5. The estimated value for no-condom infectivity λ_0 in the naive model is 18.5% lower than the estimate from the full model. Similarly the estimated value for with-condom infectivity λ_1 is about 13.5% lower than the estimate from the full model. In this case we do not know the true infectivity values, but our simulation analysis suggest that the full model is closer to the truth. All covariate parameter estimates are reasonably close to one another in both models, however the scale parameter α_2 associated with the distribution of couple-specific mean numbers of acts is about 8% higher in the naive model. This parameter is not of primary interest, but these results suggests that a naive approach may not be estimating the couple-specific act distribution very well.

We also fit an adjusted model with a variety of covariates previously shown to be associated with HIV transmission risk (Hughes et al., 2012). In addition to treatment and gender, we included viral load, male circumcision, partner HSV2 status, index HSV2 status, and partner age. Covariate parameter estimates agree reasonably well with previously published findings (Baeten et al., 2012; Hughes et al., 2012). Both treatment regimens dramatically reduce the per-act risk of HIV transmission, but the addition of FTC, although apparently beneficial, does not yield a significant reduction in risk as compared to TDF alone (Table 5.6). Male-to-female per-act infectivity is estimated to be about 50% higher than female-to-male infectivity, but after adjusting for other factors this difference is not significant Every additional year of age for the partner is associated with a 5% decline in risk. HSV2 for

	Starting Value	Full model Estimate (CI)	Naive Model Estiamte (CI)	%Difference
λ_0	0.0007	0.000868 (0.000473, 0.0014)	0.000707 (0.000421, 0.00108)	-18.5
λ_1	0.0005	0.000178 (0.000113, 0.000262)	0.000154 (0.0000959, 0.000227)	-13.5
β_1 (TDF)	-0.6	-1.35 (-1.91, -0.824)	-1.3 (-1.87, -0.763)	-3.11
β_2 (TDF/FTC)	-0.6	-1.84 (-2.56, -1.21)	-1.81 (-2.52, -1.16)	-1.89
β_3 (MTF)	0.7	0.867 (0.424, 1.32)	0.883 (0.433, 1.34)	1.84
α_1	2	1.75 (1.68, 1.82)	1.88 (1.81, 1.96)	7.68
α_2	0.5	0.458 (0.436, 0.48)	0.447 (0.427, 0.467)	-2.49
p_1	0.5	0.571 (0.566, 0.575)	-	-
p_2	0.48	0.419 (0.415, 0.424)	-	-
p_3	0.02	0.00971 (0.00924, 0.0103)	-	-

Table 5.5: Full and naive model parameter estimates, 95% credible intervals and percent difference between model estimates (PrEP data), MCMC iterations: 30,000-40,000.

either the index or partner is estimated to increase transmission risk by a non-significant amount. Surprisingly, male circumcision is estimated to reduce risk by only about 23%, and the result is not significant. Most prior studies have estimated a significant, strong effect of about 50%. The full model and the naive model give similar results for adjusted covariate effects.

	Full Model RR (95% C.I.)	Naive Model RR (95% C.I.)
TDF vs Placebo	0.266 (0.15, 0.449)	0.272 (0.154, 0.469)
TDF-FTC vs Placebo	0.166 (0.0814, 0.313)	0.165 (0.0786, 0.312)
HIV+ Male	1.45 (0.822, 2.65)	1.49 (0.844, 2.71)
Viral Load (per Log10 RNA copies/ml)	2 (1.56, 2.6)	1.98 (1.54, 2.58)
Partner Age	0.949 (0.919, 0.979)	0.95 (0.92, 0.981)
HIV- Male Circumcision	0.768 (0.386, 1.54)	0.809 (0.411, 1.59)
Partner HSV2+	1.22 (0.411, 4.38)	1.23 (0.427, 4.4)
Index HSV2+	1.13 (0.418, 3.31)	1.21 (0.451, 3.34)

Table 5.6: Estimated per-act relative risks and 95% credible intervals for covariates.

Based on equation (5.5) and full-model estimates for λ_0 and λ_1, the fraction of acts where condoms were reportedly used but where condoms were in fact not used (or failed) is approximately 22%. It should be noted that the proposed model as expressed in (5.4) assumes equivalent covariate effects for acts with and without reported condom use. This may not be true. We suspect, however, that condom use is over-reported and that the actual risk of transmission with proper condom use is close to zero. Hence covariate effects on transmission risk under proper condom use would not be estimable or meaningful. Instead we assume the true scenario is closer to the one expressed by equation (5.5) whereby covariates are associated with infectivity only in so far as they affect truly unprotected acts.

5.7 Do acts matter?

Early in the HIV epidemic Kaplan (1990) observed a "plateau" effect among transmission data whereby the probability of HIV transmission seemed to be fairly flat as a function of the number of sex acts with HIV-infected partners. This work called into question the validity of the simple transmission model, which assumes a Bernoulli model with each act having an independent and equal (within couple) probability of transmission. In fact, when we look at the observed transmission rates versus reported act counts for the Partner's PrEP data, we see something very similar to this phenomenon (Figure 5.2A). When the number of acts is larger than one, it appears that the transmission probability is uncorrelated with the number of acts.

However, it is now well known that viral load, circumcision, condom use, STIs, and other critical covariates play important roles in transmission risk. It may be the case that these other predictors of transmission risk tend to overwhelm act counts in terms of predicting the probability of transmission. For instance, imagine a simple scenario where a certain subgroup of couples has particularly high risk of transmission, and that the probability of transmission after just two sex acts is about 90%. Imagine also that all other couples have very low transmission risk. The high-risk couples are likely to transmit regardless of whether they have sex twice in an interval or five times in an interval. And so, if the proportion of high-risk couples is similar across intervals with various numbers of sex acts, the probability of transmission would tend to be about the same for all intervals. In effect, the probability of transmission would reflect the proportion of high-risk couples in each interval instead of the the number of acts. This risk-leveling effect may explain the plateau observed in Figure 5.2A.

In the framework of the full transmission model we can now think about combining the effects of sex acts with other covariates, and one way to do that is to generate the "effective"

number of acts for each interval. Since HIV infectivity is very small, we can assume the equivalence between equations (5.2) and (5.3), and instead of thinking about the covariate effects on infectivity λ as we would in equation (5.3), we think about covariate effects on the number of acts n as we would in equation (5.2). We define the effective number of acts $n_e = n * exp(\boldsymbol{x}\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the vector of covariate parameters estimates derived from the fitted model, $\boldsymbol{x}$ is the vector of observed covariates, and n is the estimated true number of acts. This new quantity n_e can be thought of as the number of acts needed by a couple with zero-valued covariates in order to produce similar transmission risk to the couple who had n acts and covariate levels $\boldsymbol{x}$.

Figure 5.2B shows the results of estimating the effective numbers of acts based on estimated latent act counts, treatment status, gender of the HIV+ partner, and viral load (centered). By plotting observed transmission frequencies by effective acts, we see a strong positive trend. (Outliers from this trend are not surprising in sparse act categories given small size of HIV infectivity.) Effective acts in this example represent numbers of acts by an untreated couple with and HIV+ female partner and an average viral load (10,000 copies/mL). On this common scale acts appear to matter very much.

Figure 5.3 shows the same pair of plots based on the reported numbers of acts (by the HIV- partner) rather than the estimated latent counts. Both the plateau effect in part A and the linear trend in part B are not as clear when based on reported acts. It seems that by estimating the latent counts we are able to generate a clearer picture of the true association between acts and transmission risk.

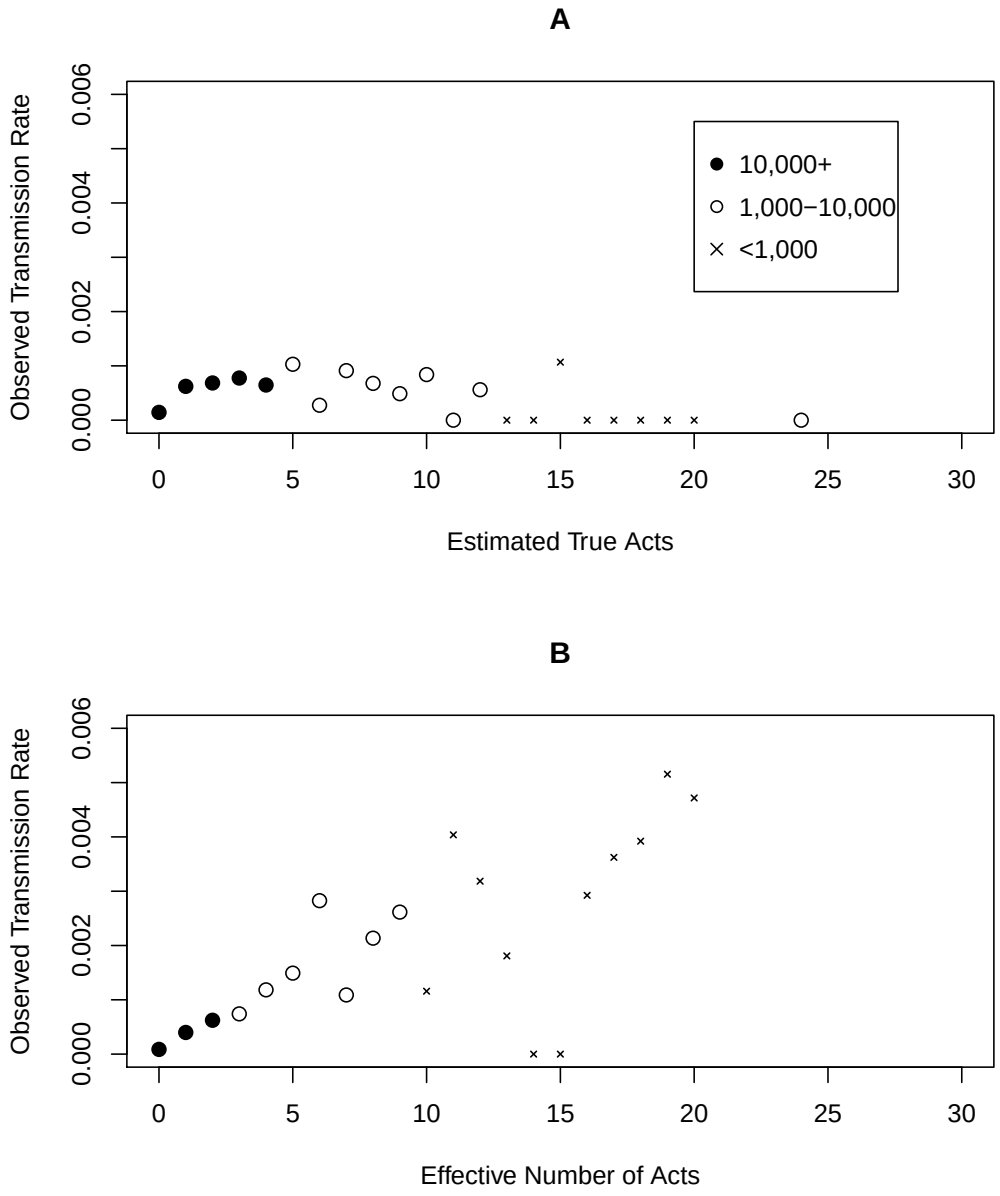

Figure 5.2: Observed transmission probability by estimated act counts and by effective act counts based on fitted model (PrEP data). Dark circles represent estimates based on 10,000 or more visits, open circles are based on 1,000-10,000 visit, and x's are based on less than 1,000 visits.

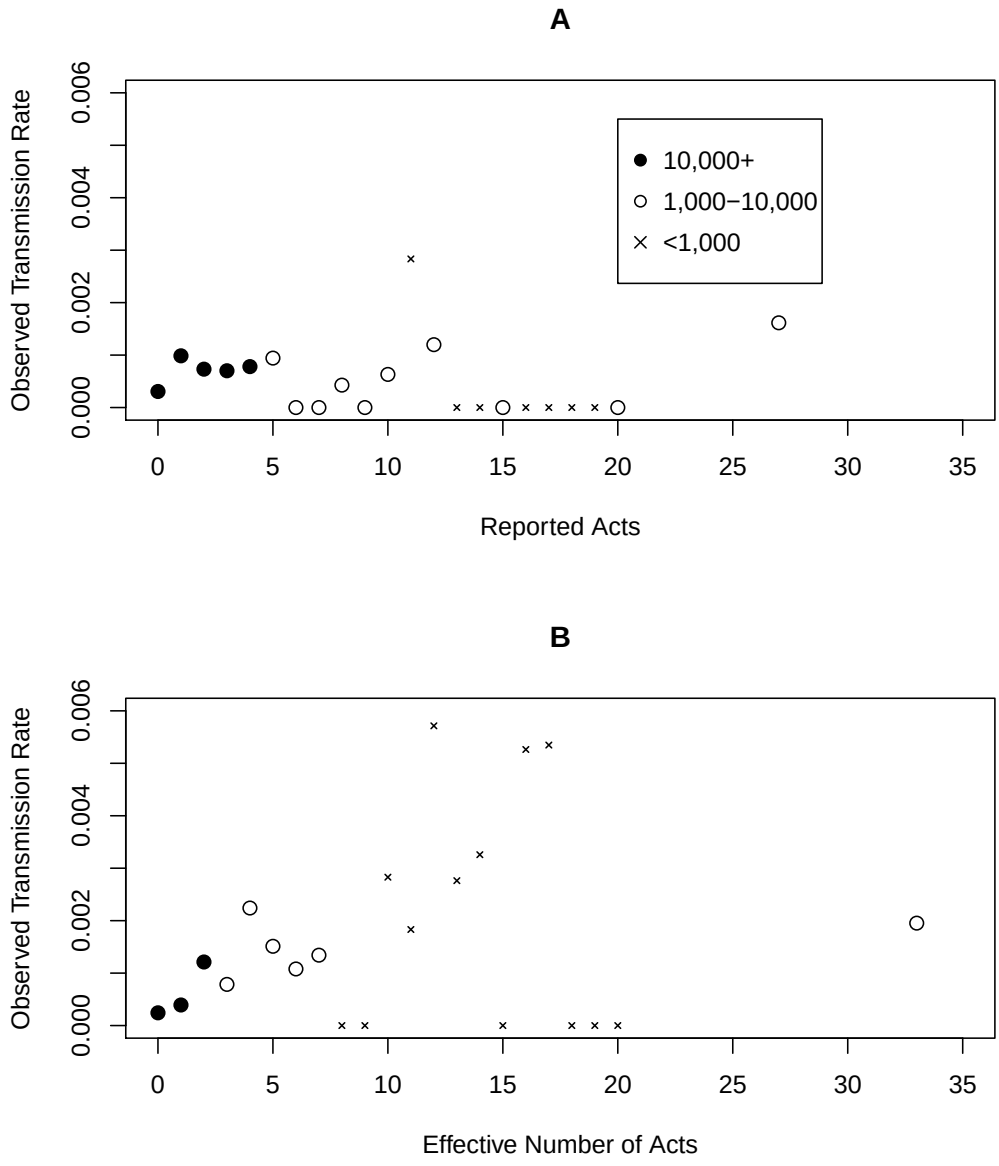

Figure 5.3: Observed transmission probability by reported acts and by effective numbers of acts based on fitted model (PrEP data). Dark circles represent estimates based on 10,000 or more visits, open circles are based on 1,000-10,000 visit, and x's are based on less than 1,000 visits.

5.8 Discussion

Simulation studies show that our proposed model can dramatically reduce per-act infectivity estimation bias in discordant partner study analyses. Our modeling of the Partners PrEP data suggests that HIV1 per-act infectivity may be nearly 25% higher than would be estimated using naive models.

These results suggest that collecting independent, replicate reports of sexual history in partner studies is worth the added effort. In most partner studies both partners will be monitored at regular intervals anyway, either assessing treatment compliance, monitoring viral load, or testing for transmission. The addition of a brief sexual history survey to a pre-existing clinical visit should generally be a minimal added burden. In fact, our results show that only partial replication is necessary. In the Partners PrEP study we analyzed here, HIV positive partners were seen and interviewed only once for every three visits by the HIV negative partner, and still the added benefit of collecting that data was substantial.

An important assumption that we make in our simulation procedure is that the misreporting of sex acts is itself unbiased. If, in fact, the reported number of acts is higher on average than the true number of acts, then our estimate of infectivity would tend to be too low. The opposite would be true for under reporting of sexual activity. In general, we do not have reason to suspect an over or under reporting of sexual activity, although Clark and Wallin (1964) provide compelling evidence that biased reporting can happen in certain subgroups. For example, we might expect under-reporting when subjects are queried about sex acts following abstinence counseling, but we don't attempt to model those issues here.

Our latent variable model assumes that reported acts by opposing members of a partnership are independent conditional on the true number of acts. Brenner (1992) warns that if this is not the case, latent-class methods can only correct for a portion of misclassifica-

tion bias. In dual-reporting partner studies each partner is interviewed separately, often on different days, and so we do not expect misreporting to be correlated. However, if partners were to discuss their sexual activity prior to the interview, it is possible that this condition could be violated.

Another assumption is that the mis-reporting of sex acts, when present, roughly follows a Poisson distribution. Our simulations suggest that in the case where mis-reporting is overdispersed as compared to Poisson measurement error, the proposed model still gives unbiased estimates. This may not be the case for other forms of mis-reporting, but it is nevertheless reassuring. Measurement error models could be better informed by a validity study of sexual history reporting, but such studies are difficult to conduct and even a high quality study may not be able to accurately quantify the true number of sex acts for couples over time. An important advantage of our model is that a wide variety of mis-reporting distributions can be used, depending on the context, including mixtures of distributions.

An important innovation in our model is its ability to accommodate non-overlapping time intervals. In cases where study partners are interviewed at nearly the same time this feature may not make a large difference. However, when partner interviews are sporadic and frequently off schedule, the ability to accurately model mis-timed reports using latent variables becomes important. This strategy introduces a new level of flexibility for study design, whereby the importance of maintaining a strict interview schedule becomes less important.

In summary, the model we introduce in this paper is a novel, flexible approach to modeling the inherently complex data associated with partner studies. By exploiting the dual-reporting of sexual history, this model is able to reduce, if not eliminate, the bias associated with estimating per-act infectivity in the presence of act mis-reporting.

5.9 Appendix A - Full conditional distribution and Gibbs sampling procedure.

The full conditional distributions of each model parameter, not including the normalizing constants, are as follows:

$$
\begin{aligned}
p(\lambda_0|\boldsymbol{y},\boldsymbol{n},\boldsymbol{n1},\lambda_1,\boldsymbol{x},\boldsymbol{\beta}) &\propto p(\boldsymbol{y}|\lambda_0,\lambda_1,\boldsymbol{n},\boldsymbol{n1},\boldsymbol{x},\boldsymbol{\beta})\pi(\lambda_0) \qquad (5.33)\\
&= \prod_{i=1}^{N}\prod_{k=1}^{TV_i} p(y_{ik}|\lambda_0,\lambda_1,n'_{ik},n1'_{ik},x_{ik},\boldsymbol{\beta})\pi(\lambda_0)\\
&\propto \prod_{i=1}^{N}\prod_{k=1}^{TV_i} \left[1-(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}}\\
&\times \left[(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}}\\
&\times \lambda_0^{\phi_1-1}(1-\lambda_0)^{\phi_2-1}
\end{aligned}
$$

$$
\begin{aligned}
p(\lambda_1|\boldsymbol{y},\boldsymbol{n},\boldsymbol{n1},\lambda_0,\boldsymbol{x},\boldsymbol{\beta}) &\propto p(\boldsymbol{y}|\lambda_0,\lambda_1,\boldsymbol{n},\boldsymbol{n1},\boldsymbol{x},\boldsymbol{\beta})\pi(\lambda_1) \qquad (5.34)\\
&= \prod_{i=1}^{N}\prod_{k=1}^{TV_i} p(y_{ik}|\lambda_0,\lambda_1,n'_{ik},n1'_{ik},x_{ik},\boldsymbol{\beta})\pi(\lambda_1)\\
&\propto \prod_{i=1}^{N}\prod_{k=1}^{TV_i} \left[1-(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}}\\
&\times \left[(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}}\\
&\times \lambda_1^{\tau_1-1}(1-\lambda_1)^{\tau_2-1}
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{n},\boldsymbol{n1},\lambda_0,\lambda_1,\boldsymbol{x}) &\propto p(\boldsymbol{y}|\lambda_0,\lambda_1,\boldsymbol{n},\boldsymbol{n1},\boldsymbol{x},\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \qquad (5.35)\\
&= \prod_{i=1}^{N}\prod_{k=1}^{TV_i} p(y_{ik}|\lambda_0,\lambda_1,n'_{ik},n1'_{ik},x_{ik},\boldsymbol{\beta})\pi(\boldsymbol{\beta})\\
&\propto \prod_{i=1}^{N}\prod_{k=1}^{TV_i} \left[1-(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}}\\
&\times \left[(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}}\\
&\times exp(-\frac{1}{2}\boldsymbol{\beta}^T\Sigma^{-1}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\psi}|\boldsymbol{n},\boldsymbol{t},\alpha_1,\alpha_2) &\propto p(\boldsymbol{n}|\boldsymbol{\psi},\boldsymbol{t})\pi(\boldsymbol{\psi}|\alpha_1,\alpha_2) \qquad (5.36)\\
&= \prod_{i=1}^{N}\prod_{j=1}^{J_i} p(n_{ij}|\psi_i,t_{ij},t_{i(j-1)})\pi(\psi_i|\alpha_1,\alpha_2)\\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{J_i} \frac{1}{n_{ij}!}(\psi_i(t_{ij}-t_{i(j-1)})/28)^{n_{ij}} exp(-\psi_i(t_{ij}-t_{i(j-1)})/28)\\
&\times \prod_{i=1}^{N}\psi_i^{\alpha_1-1}exp(-\alpha_2\psi_i)
\end{aligned}
$$

$$
\begin{aligned}
p(\alpha_1|\boldsymbol{\psi},\alpha_2) &\propto \pi(\boldsymbol{\psi}|\alpha_1,\alpha_2)\pi(\alpha_1) \qquad (5.37)\\
&= \prod_{i=1}^{N}\pi(\psi_i|\alpha_1,\alpha_2)\pi(\alpha_1))\\
&\propto \prod_{i=1}^{N}\alpha_2^{\alpha_1}\psi_i^{\alpha_1-1}exp(-\alpha_2\psi_i)\times\alpha_1^{\delta_1-1}exp(-\delta_2\alpha_1)\\
p(\alpha_2|\boldsymbol{\psi},\alpha_1) &\propto \pi(\boldsymbol{\psi}|\alpha_1,\alpha_2)\pi(\alpha_2) \qquad (5.38)\\
&= \prod_{i=1}^{N}\pi(\psi_i|\alpha_1,\alpha_2)\pi(\alpha_2))\\
&\propto \prod_{i=1}^{N}\alpha_2^{\alpha_1}\psi_i^{\alpha_1-1}exp(-\alpha_2\psi_i)\times\alpha_2^{\omega_1-1}exp(-\omega_2\alpha_2)\\
&\propto Gamma(N\alpha_1+\omega_1,\sum_{i=1}^{N}\psi_i+\omega_2)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{p.condom}|\boldsymbol{n},\boldsymbol{n1},\nu_1,\nu_2) &\propto p(\boldsymbol{n1}|\boldsymbol{n},\boldsymbol{p.condom})\pi(\boldsymbol{p.condom}|\nu_1,\nu_2) \qquad (5.39)\\
&= \prod_{i=1}^{N}\prod_{j=1}^{J_i} p(n1_{ij}|n_{ij},p.condom_i)\pi(p.condom_i|\nu_1,\nu_2)\\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{J_i} p.condom_i^{n1_{ij}}(1-p.condom_i)^{n_{ij}-n1_{ij}}\\
&\times \prod_{i=1}^{N} p.condom_i^{\nu_1-1}(1-p.condom_i)^{\nu_2-1}\\
p.condom_i &\sim Beta(\sum_{j=1}^{J_i} n1_{ij}+\nu_1, \sum_{j=1}^{J_i}(n_{ij}-n1_{ij})+\nu_2)
\end{aligned}
$$

$$
\begin{aligned}
p(\nu_1|\boldsymbol{p.condom},\nu_2) &\propto \pi(\boldsymbol{p.condom}|\nu_1,\nu_2)\pi(\nu_1) \qquad (5.40)\\
&= \prod_{i=1}^{N}\pi(p.condom_i|\nu_1,\nu_2)\times\pi(\nu_1))\\
&\propto \prod_{i=1}^{N} p.condom_i^{\nu_1-1}(1-p.condom_i)^{\nu_2-1}\times\nu_1^{\mu_1-1}(1-\nu_1)^{\mu_2-1}\\
p(\nu_2|\boldsymbol{p.condom},\nu_1) &\propto \pi(\boldsymbol{p.condom}|\nu_1,\nu_2)\pi(\nu_2) \qquad (5.41)\\
&= \prod_{i=1}^{N}\pi(p.condom_i|\nu_1,\nu_2)\times\pi(\nu_2))\\
&\propto \prod_{i=1}^{N} p.condom_i^{\nu_1-1}(1-p.condom_i)^{\nu_2-1}\times\nu_2^{\kappa_1-1}(1-\nu_2)^{\kappa_2-1}
\end{aligned}
$$

(5.42)

$$
\begin{aligned}
& p(\boldsymbol{n}|\boldsymbol{y}, \boldsymbol{n1}, \lambda_0, \lambda_1, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{t}, \boldsymbol{m0}, \boldsymbol{p.condom}) \\
\propto\ & p(\boldsymbol{y}|\lambda_0, \lambda_1, \boldsymbol{n}, \boldsymbol{n1}, \boldsymbol{x}, \boldsymbol{\beta}) p(\boldsymbol{n}|\boldsymbol{\psi}, \boldsymbol{t}) p(\boldsymbol{n}1|\boldsymbol{n}, \boldsymbol{p.condom}) p(\boldsymbol{m0}|\boldsymbol{n}, \boldsymbol{n1}) \qquad (5.43) \\
=\ & \prod_{i=1}^{N} \prod_{k=1}^{TV_i} p(y_{ik}|\lambda_0, \lambda_1, n'_{ik}, n1'_{ik}, \boldsymbol{x}_{ik}, \boldsymbol{\beta}) \\
\times\ & \prod_{j=1}^{J_i} p(n_{ij}|\psi_i, t_{ij}, t_{i(j-1)}) p(n1_{ij}|n_{ij}, p.condom_i) \\
\times\ & \prod_{l=1}^{2} \prod_{v=1}^{V_{il}} p(m0_{ivl}|n^*_{ivl}, n1^*_{ivl}) \\
\propto\ & \prod_{i=1}^{N} \prod_{k=1}^{TV_i} \left[1 - (1 - \lambda_0 exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta}))^{n0'_{ik}} (1 - \lambda_1 exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}} \\
\times\ & \left[(1 - \lambda_0 exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta}))^{n0'_{ik}} (1 - \lambda_1 exp(\boldsymbol{x}_{ik}^T \boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}} \\
\times\ & \prod_{j=1}^{J_i} \frac{1}{n_{ij}!} (\psi_i (t_{ij} - t_{i(j-1)})/28)^{n_{ij}} exp(-\psi_i (t_{ij} - t_{i(j-1)})/28) \\
\times\ & \prod_{j=1}^{J_i} \binom{n_{ij}}{n1_{ij}} p.condom_i^{n1_{ij}} (1 - p.condom_i)^{n_{ij} - n1_{ij}} \\
\times\ & \prod_{l=1}^{2} \prod_{v=1}^{V_{il}} n0_{ivl}^{*m0_{ivl}} exp(-n0^*_{ivl})
\end{aligned}
$$

$$
\begin{aligned}
& p(\boldsymbol{n1}|\boldsymbol{y},\boldsymbol{n},\lambda_0,\lambda_1,\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{m1},\boldsymbol{m0},\boldsymbol{p.condom}) \\
\propto\ & p(\boldsymbol{y}|\lambda_0,\lambda_1,\boldsymbol{n},\boldsymbol{n1},\boldsymbol{x},\boldsymbol{\beta})p(\boldsymbol{n1}|\boldsymbol{n},\boldsymbol{p.condom})p(\boldsymbol{m1}|\boldsymbol{n1})p(\boldsymbol{m0}|\boldsymbol{n},\boldsymbol{n1}) \qquad (5.44)\\
=\ & \prod_{i=1}^{N}\prod_{k=1}^{TV_i} p(y_{ik}|\lambda_0,\lambda_1,n'_{ik},n1'_{ik},\boldsymbol{x}_{ik},\boldsymbol{\beta}) \\
\times\ & \prod_{j=1}^{J_i} p(n1_{ij}|n_{ij},p.condom_i) \\
\times\ & \prod_{l=1}^{2}\prod_{v=1}^{V_{il}} p(m0_{ivl}|n^*_{ivl},n1^*_{ivl})p(m1_{ivl}|n1^*_{ivl}) \\
\propto\ & \prod_{i=1}^{N}\prod_{k=1}^{TV_i} \left[1-(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{y_{ik}} \\
\times\ & \left[(1-\lambda_0 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n0'_{ik}}(1-\lambda_1 exp(\boldsymbol{x}_{ik}^T\boldsymbol{\beta}))^{n1'_{ik}}\right]^{1-y_{ik}} \\
\times\ & \prod_{j=1}^{J_i}\binom{n_{ij}}{n1_{ij}} p.condom_i^{n1_{ij}}(1-p.condom_i)^{n_{ij}-n1_{ij}} \\
\times\ & \prod_{l=1}^{2}\prod_{v=1}^{V_{il}} n1_{ivl}^{*m1_{ivl}} exp(-n1^*_{ivl}) \times n0_{ivl}^{*m0_{ivl}} exp(-n0^*_{ivl})
\end{aligned}
$$

Only α_2 and $p.condom_i$ have conjugate distributions and can be sampled directly. All others will be sampled using a Metropolis-Hastings (or similar) approach.

To generate a sample from the joint posterior distribution of the parameters and latent variables conditional on observed data, select staring values

$$
(\lambda_0^{(0)},\lambda_1^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{\psi}^{(0)},\boldsymbol{p.condom}^{(0)},\alpha_1^{(0)},\alpha_2^{(0)},\nu_1^{(0)},\nu_2^{(0)},\boldsymbol{n}^{(0)},\boldsymbol{n1}^{(0)}). \qquad (5.45)
$$

Then, using Gibbs sampling, generate a sequence of values from the conditionals where each parameter depends on the current (nth) and previous $((n-1)th)$ iteration as follows:

$$\lambda_0^{(n)} \sim p(\lambda_0^{(n)}|\boldsymbol{y}, \boldsymbol{n}^{(n-1)}, \boldsymbol{n1}^{(n-1)}, \lambda_1^{(n-1)}, \boldsymbol{x}, \boldsymbol{\beta}^{(n-1)}) \quad (5.46)$$

$$\lambda_1^{(n)} \sim p(\lambda_1^{(n)}|\boldsymbol{y}, \boldsymbol{n}^{(n-1)}, \boldsymbol{n1}^{(n-1)}, \lambda_0^{(n)}, \boldsymbol{x}, \boldsymbol{\beta}^{(n-1)}) \quad (5.47)$$

$$\boldsymbol{\beta}^{(n)} \sim p(\boldsymbol{\beta}^{(n)}|\boldsymbol{y}, \boldsymbol{n}^{(n-1)}, \boldsymbol{n1}^{(n-1)}, \lambda_0^{(n)}, \lambda_1^{(n)}, \boldsymbol{x}) \quad (5.48)$$

$$\boldsymbol{\psi}^{(n)} \sim p(\boldsymbol{\psi}^{(n)}|\boldsymbol{n}^{(n-1)}, \boldsymbol{t}, \alpha_1^{(n-1)}, \alpha_2^{(n-1)}) \quad (5.49)$$

$$\alpha_1^{(n)} \sim p(\alpha_1^{(n)}|\boldsymbol{\psi}^{(n)}, \alpha_2^{(n-1)}) \quad (5.50)$$

$$\alpha_2^{(n)} \sim p(\alpha_2^{(n)}|\boldsymbol{\psi}^{(n)}, \alpha_1^{(n)}) \quad (5.51)$$

$$\boldsymbol{p.condom}^{(n)} \sim p(\boldsymbol{p.condom}^{(n)}|\boldsymbol{n}^{(n-1)}, \boldsymbol{n1}^{(n-1)}, \nu_1^{(n-1)}, \nu_2^{(n-1)}) \quad (5.52)$$

$$\nu_1^{(n)} \sim p(\nu_1^{(n)}|\boldsymbol{p.condom}^{(n)}, \nu_2^{(n-1)}) \quad (5.53)$$

$$\nu_2^{(n)} \sim p(\nu_1^{(n)}|\boldsymbol{p.condom}^{(n)}, \nu_1^{(n)}) \quad (5.54)$$

$$\boldsymbol{n}^{(n)} \sim p(\boldsymbol{n}^{(n)}|\boldsymbol{y}, \boldsymbol{n1}^{(n-1)}, \lambda_0^{(n)}, \lambda_1^{(n)}, \boldsymbol{x}, \boldsymbol{\beta}^{(n)}, \boldsymbol{t}, \boldsymbol{m0}, \boldsymbol{p.condom}^{(n)}) \quad (5.55)$$

$$\boldsymbol{n1}^{(n)} \sim p(\boldsymbol{n1}^{(n)}|\boldsymbol{y}, \boldsymbol{n}^{(n)}, \lambda_0^{(n)}, \lambda_1^{(n)}, \boldsymbol{x}, \boldsymbol{\beta}^{(n)}, \boldsymbol{m1}, \boldsymbol{m0}, \boldsymbol{p.condom}^{(n)}) \quad (5.56)$$

Sampling continues until it appears that the simulated observations represent a sample from the stationary distribution of all parameters and latent variables, conditional on the observed data.

5.10 Appendix B - Convergence Diagnostics for Simulations

Trace plots for a single simulation (Figures 5.4 and 5.5) suggest that the posterior sampling distributions for the primary parameters are almost immediately stable, whereas the distributions for the nuisance parameters stabilize after approximately 5,000 iterations. If the first 5,000 iterations are dropped (Figures 5.6 and 5.7), the resulting trace plots suggest that sampling the distributions have stabilized for all parameters. Although the trace plots for $\boldsymbol{p}$ show a notable degree of autocorrelation, they all appear to be fairly stable and have consistent means. Table 5.7 contains empirical means and 95% credible intervals based on three different segments of the MCMC chain. Observed means change slightly after the first 5,000 iterations of the MCMC sampler, but very little after that. Figure 5.8 shows plots of the prior distributions for each parameter, and the estimated posterior distributions based on the MCMC sample for the first simulated dataset. Posteriors do not appear to be unduly influenced by the priors or the starting values. Multiple chains with different starting values were also run, and convergence was evident within 1,000 iterations (plots not shown.)

	MCMC Sample Range		
Parameter	1-5,000	5,000-10,000	10,000-15,000
λ_0	0.00205 (0.00131, 0.003)	0.00206 (0.00132, 0.00298)	0.00206 (0.00132, 0.00297)
λ_1	0.000395 (0.000255, 0.000571)	0.000412 (0.00027, 0.000585)	0.000416 (0.000266, 0.000594)
β_1	-0.922 (-1.4, -0.47)	-0.918 (-1.37, -0.474)	-0.916 (-1.38, -0.463)
β_2	1.32 (0.898, 1.78)	1.31 (0.871, 1.75)	1.31 (0.877, 1.76)
α_1	2.05 (1.87, 2.37)	1.97 (1.83, 2.12)	1.96 (1.83, 2.1)
α_2	0.397 (0.362, 0.439)	0.392 (0.361, 0.424)	0.395 (0.365, 0.426)
p_1	0.473 (0.428, 0.487)	0.481 (0.472, 0.489)	0.481 (0.472, 0.49)
p_2	0.521 (0.506, 0.566)	0.513 (0.505, 0.522)	0.513 (0.504, 0.522)
p_3	0.00599 (0.00538, 0.00662)	0.00593 (0.00542, 0.0065)	0.00577 (0.00512, 0.00638)

Table 5.7: Posterior means and 95% Credible Intervals for various sections of the MCMC sample (simulated data).

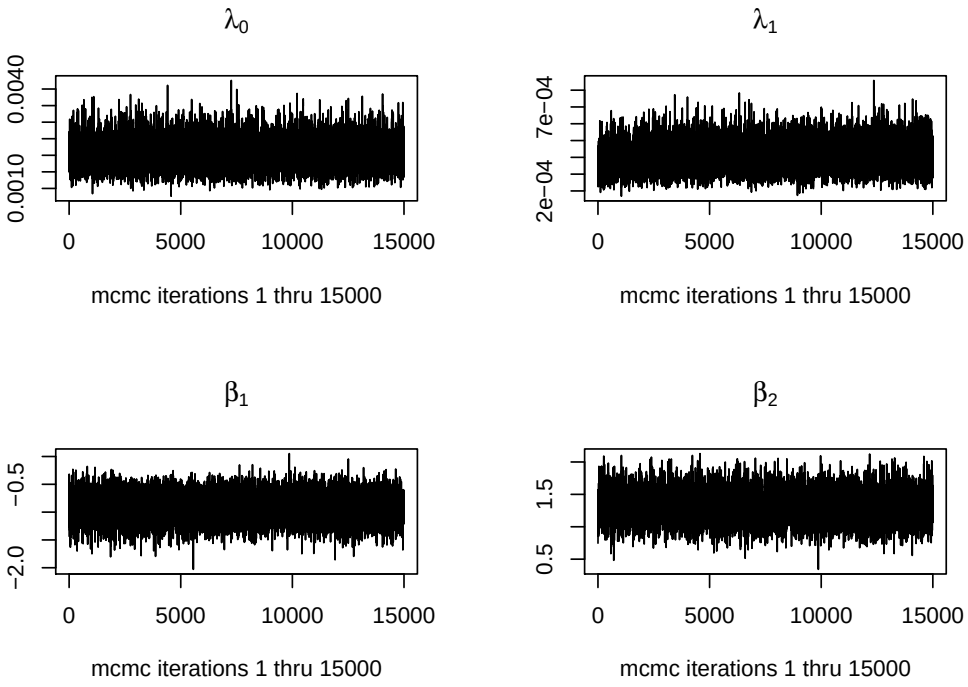

Figure 5.4: Complete traceplots for MCMC samples for the simulated dataset, primary parameters.

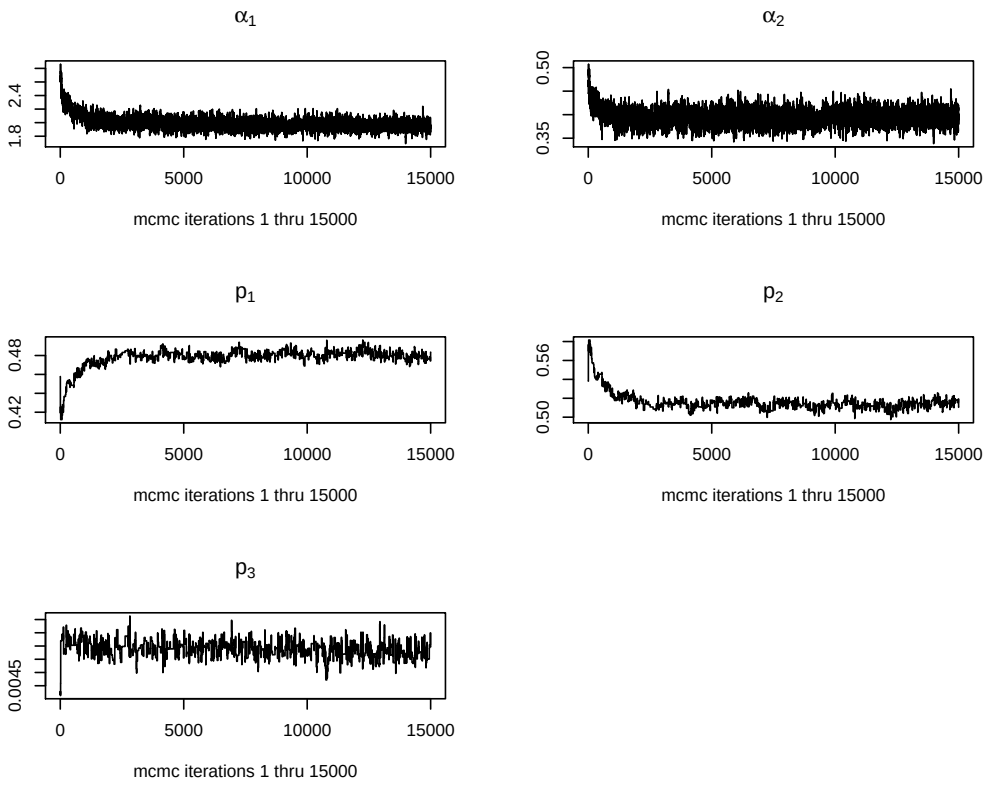

Figure 5.5: Complete traceplots for MCMC samples for the simulated dataset, secondary parameters.

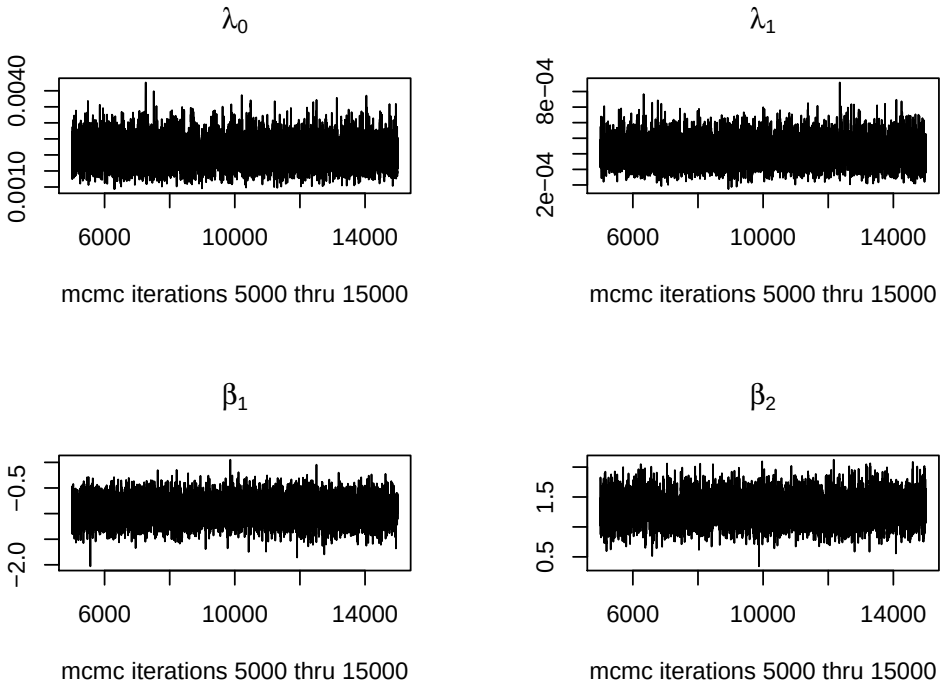

Figure 5.6: Trace plots for MCMC iterations 5,000 to 15,000 for the simulated dataset, primary parameters.

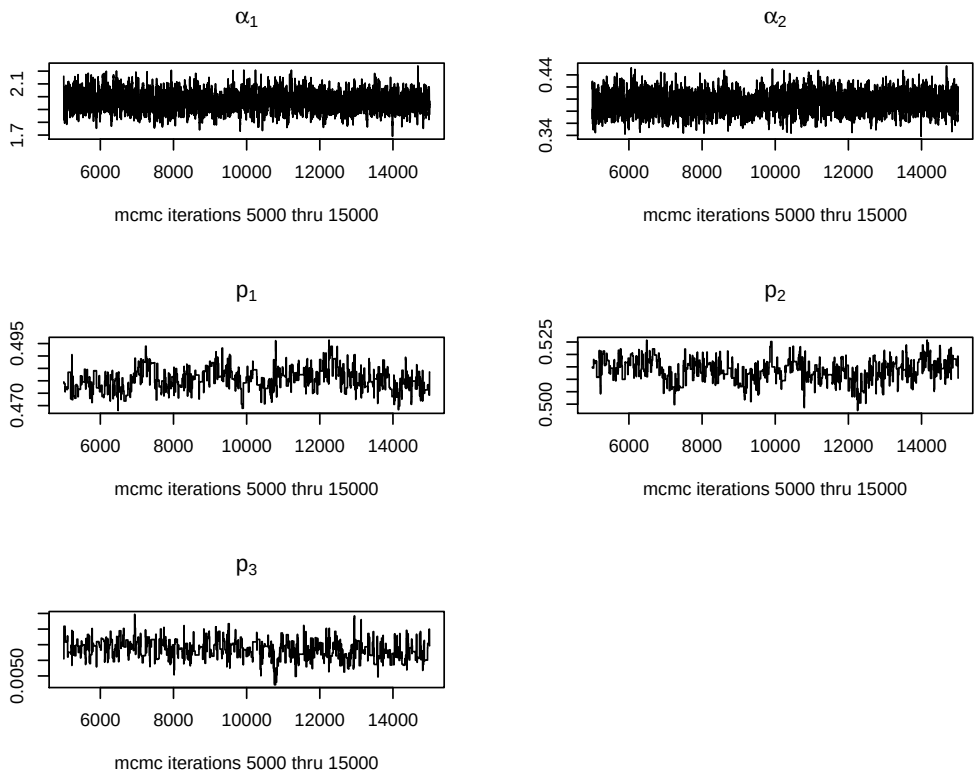

Figure 5.7: Trace plots for MCMC iterations 5,000 to 15,000 for the simulated dataset, secondary parameters.

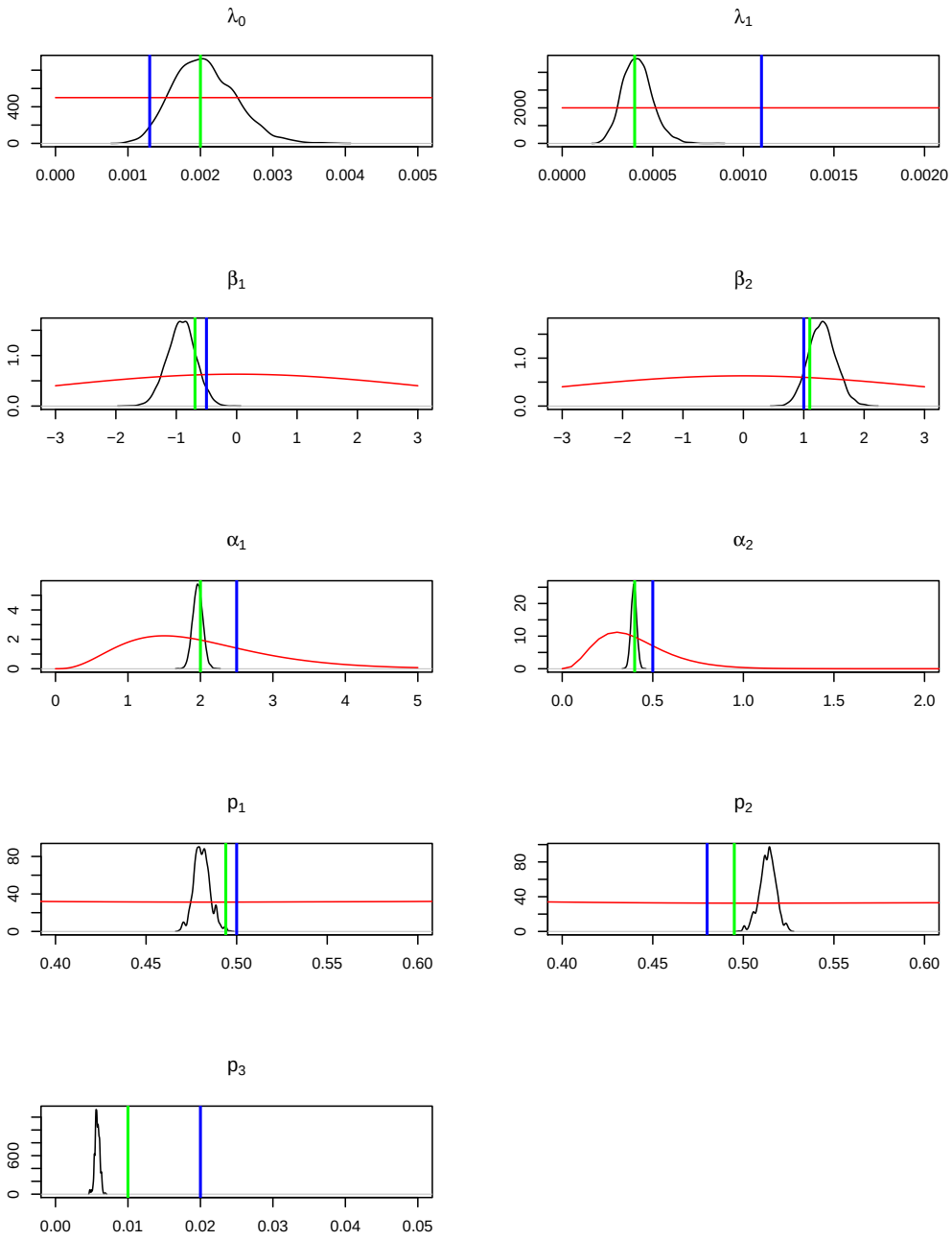

Figure 5.8: Prior distributions (red), starting values (blue), true values (green) and posterior distributions (black) for each parameter (simulated data).

5.11 Appendix C - Convergence Diagnostics for PrEP Data Analysis

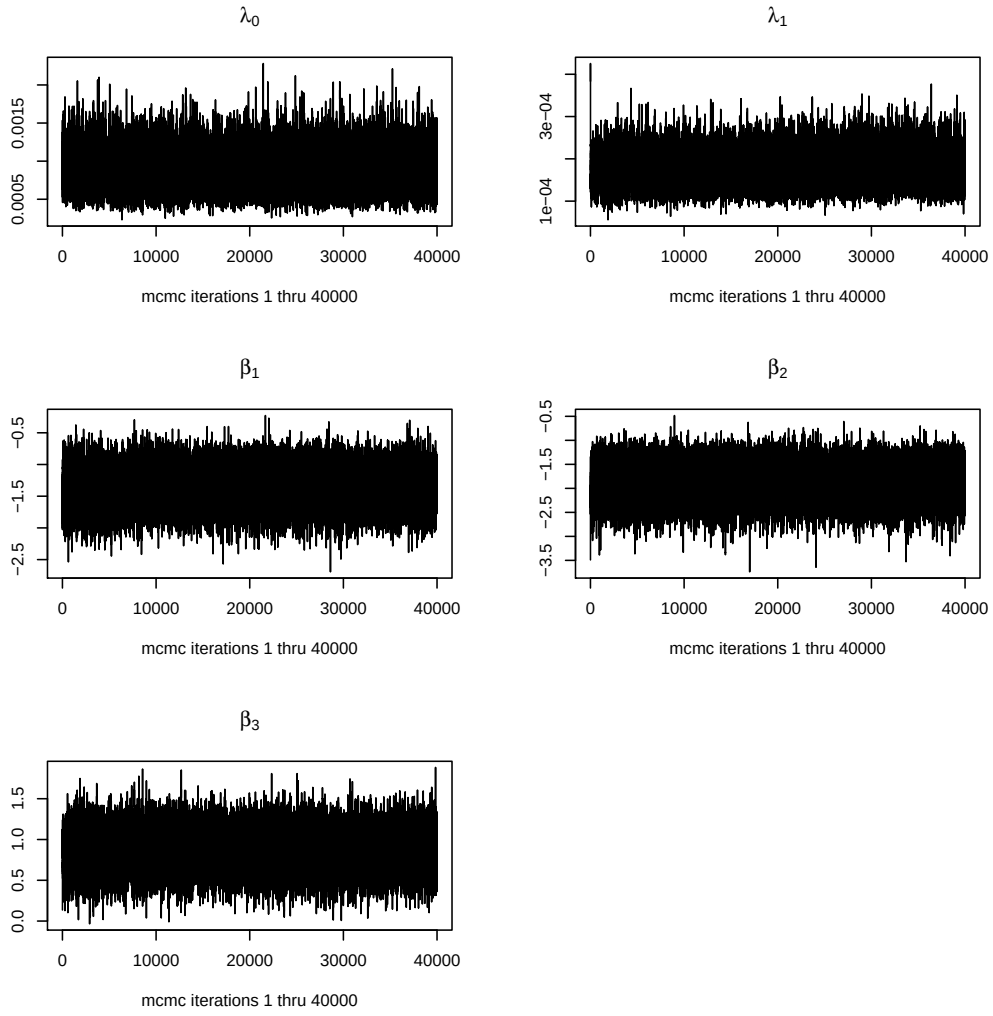

Figure 5.9: Complete traceplots of all MCMC samples for PrEP primary parameters.

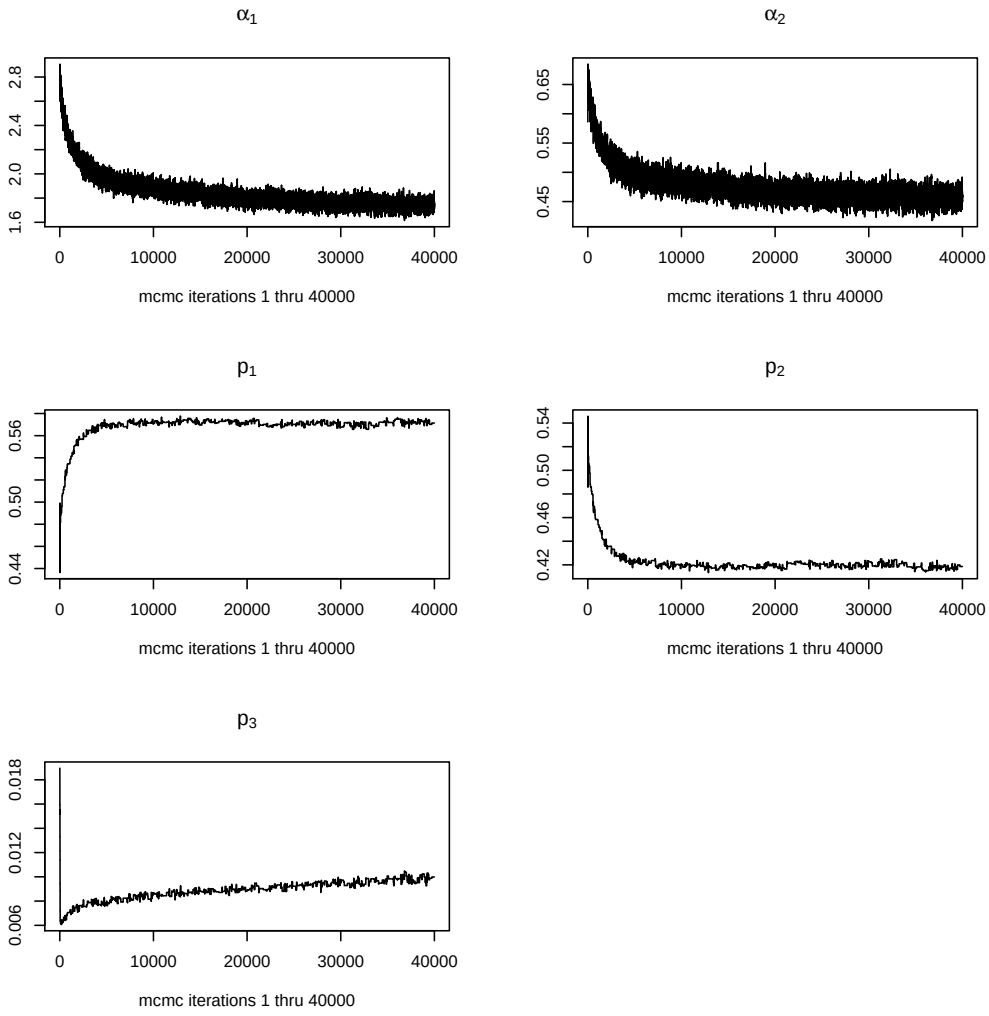

Figure 5.10: Complete traceplots of all MCMC samples for PrEP secondary parameters.

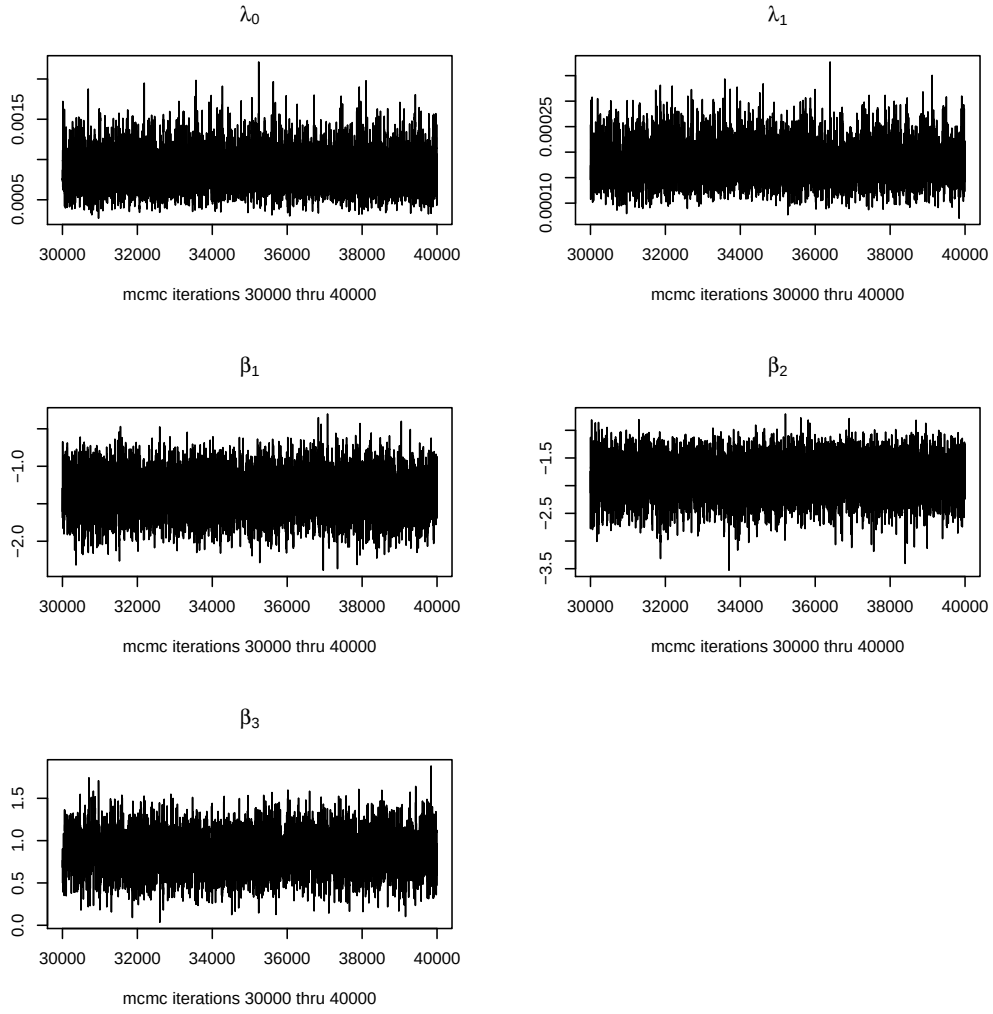

Figure 5.11: Trace plots for MCMC iterations 80,000 to 100,000 for PrEP primary parameters.

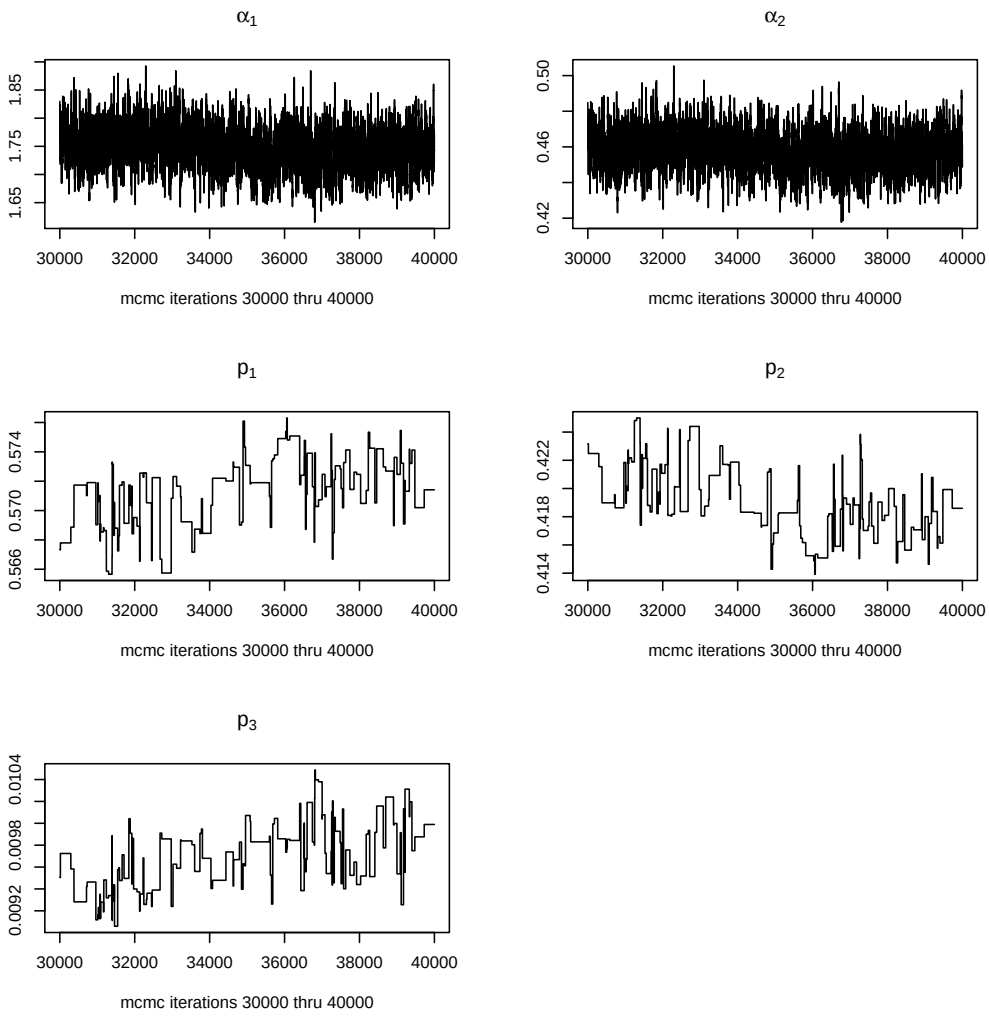

Figure 5.12: Trace plots for MCMC iterations 80,000 to 100,000 for PrEP secondary parameters.

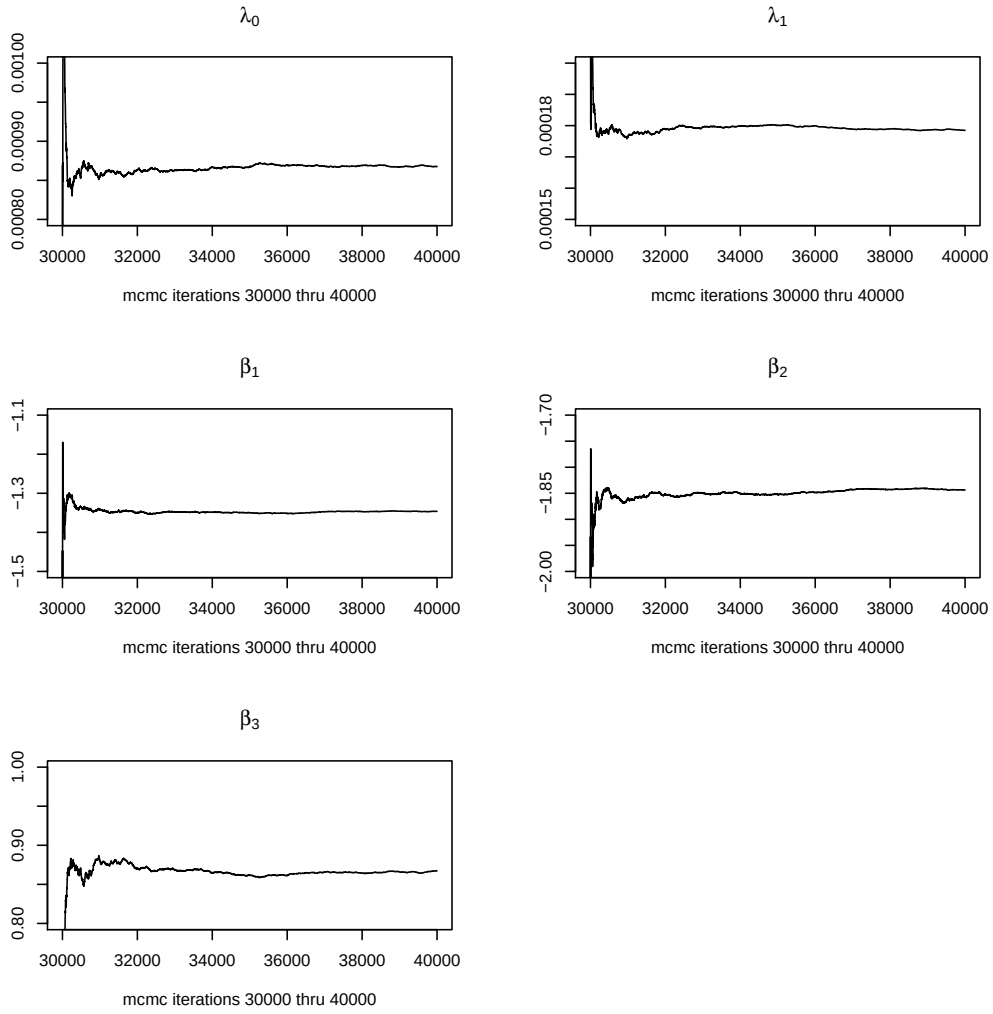

Figure 5.13: Cumulative mean plots for MCMC iterations 80,000 to 100,000 for PrEP primary parameters.

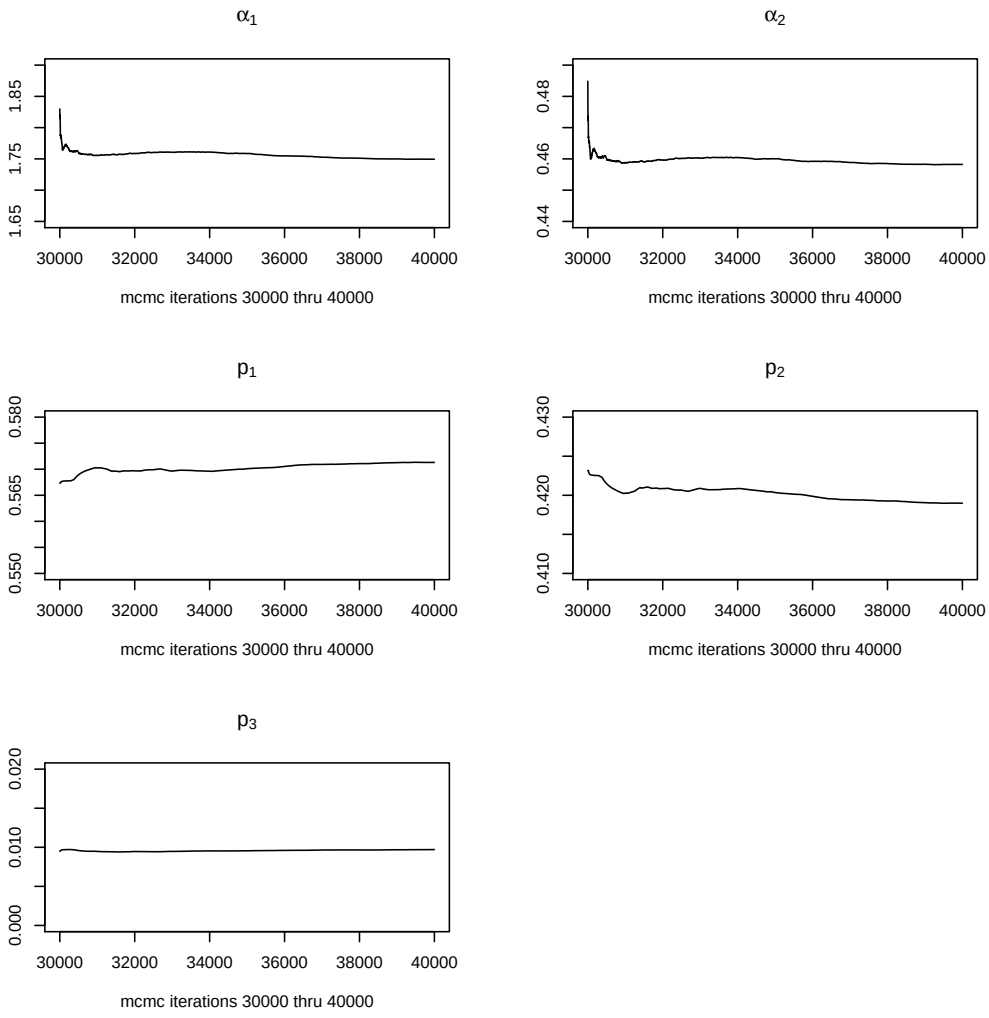

Figure 5.14: Cumulative mean plots for MCMC iterations 80,000 to 100,000 for PrEP secondary parameters.

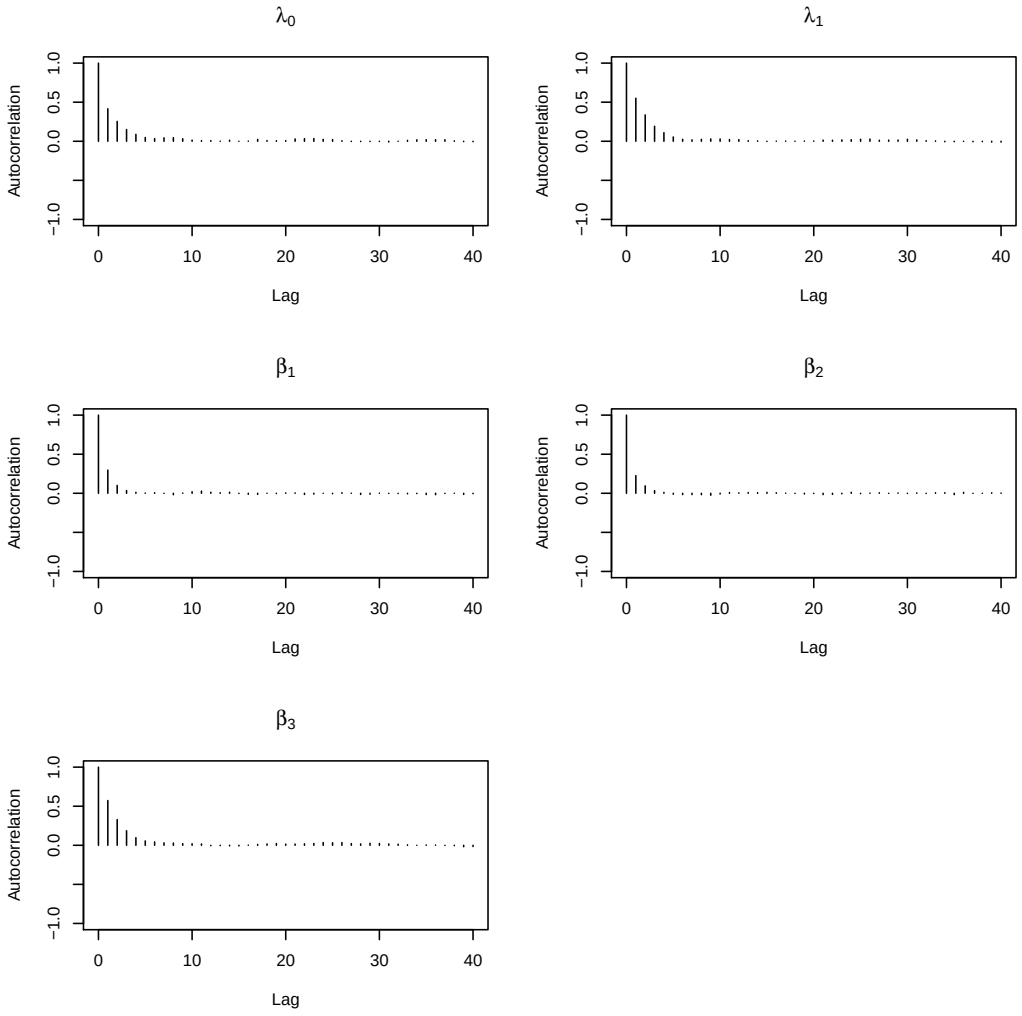

Figure 5.15: Autocorrelation plots for MCMC iterations 80,000 to 100,000 for PrEP primary parameters.

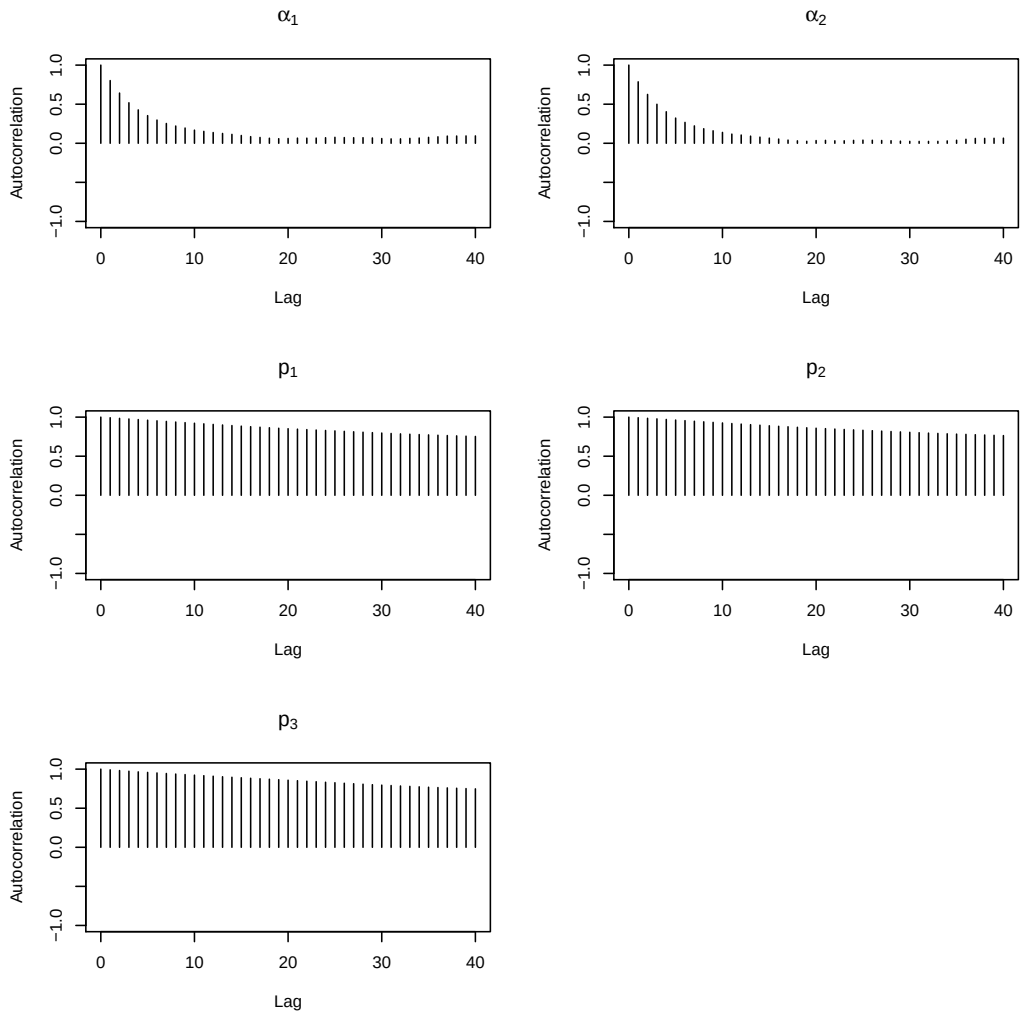

Figure 5.16: Autocorrelation plots for MCMC iterations 80,000 to 100,000 for PrEP secondary parameters.

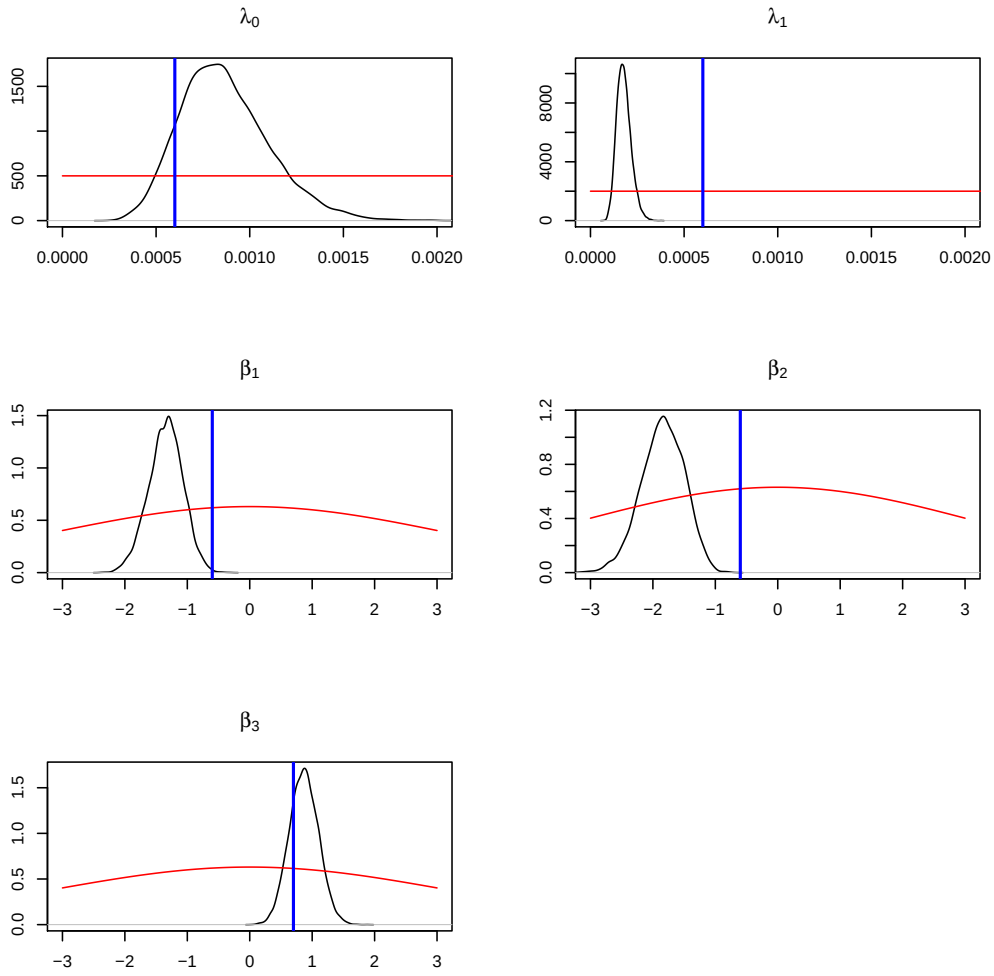

Figure 5.17: Prior distributions (red), starting values (blue) and posterior distributions (black) for PrEP analysis primary parameters.

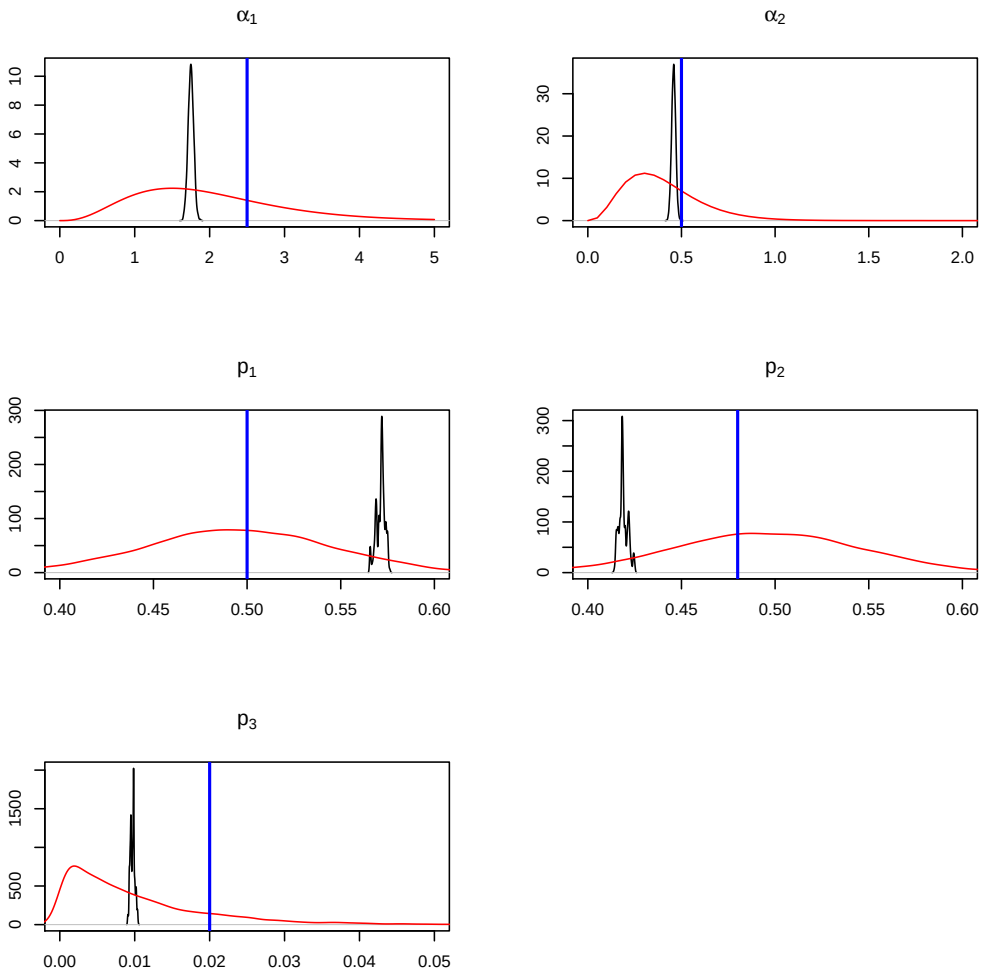

Figure 5.18: Prior distributions (red), starting values (blue) and posterior distributions (black) for PrEP analysis secondary parameters.

Chapter 6

FUTURE WORK

6.1 Pooled testing methods

Although the method proposed in Chapter 3 worked reasonably well for simulated datasets, it did not perform well when applied to real viral-load data collected from pooled specimens. One reason for this poor performance is likely the high level of apparent measurement error in the HIV RNA assay data. As discussed in Chapter 3, several of the observed row and columns pool values in Tables 3.13-3.16 would not be possible unless there had been either transcription errors or unless the RNA testing device tended to produce widely varying viral load measurements for aliquots of the same blood sample. If the problem is in fact measurement variability, then we need to better understand how much variability is acceptable and how much variability would preclude the use of our proposed pooling methods.

A natural place to start would be a simulation study that tests a broad range of measurement error variances and identifies the point at which measurement error eliminates efficiency gained by using the pooled-testing algorithm. The magnitude of measurement error variance is primarily important in relation the size of the difference between mean viral load levels for non-failures and failures, and in relation to the viral-load variances of non-failures and failures. We therefore would need to assess measurement error as a percent of difference in mean viral load or as a percent of viral load variability among non-failures and failures.

In addition, a better understanding of HIV RNA viral load testing variability is necessary in order to understand whether pooling methods can be effective. Brambilla et al.

(1999) provide valuable results regarding assay variability and biological variability for the Roche-Monitor RT-PCR assays. These analyses needs to be repeated for current assay technology and for a variety of applied settings. It may be possible to control assay variability in carefully monitored settings with the best technology, but standard laboratory procedures in relevant geographical regions may not be rigorous enough to facilitate pooling methods.

Another area of further investigation is attempting to streamline the Bayes-MCMC sampling procedure. In its current form the modeling algorithm requires several hours to complete a single round. For practical testing purposes this may not be a problem, but in the context of simulation studies that are testing dozens of parameter configurations with hundreds of repetitions, this time requirement can be prohibitive. There may be a way to restrict the MCMC sampling space in such a way that convergence is achieved much more quickly.

6.2 Estimating per-act infectivity

The proposed latent-variable model generated interesting results for the Partners PrEP study, as reported in Chapter 5, and we would like to apply this model to other discordant-couple studies, including the Partners HSV/HIV study (Celum et al., 2010). It would be valuable to have a second study with which to verify that the proposed model results in higher estimates of HIV infectivity than naive models. It also may be possible to combine data from the PrEP and HSV/HIV studies to generate robust estimates of infectivity and covariate parameters.

Another area to explore is the problem of viral-load measurement error. It is well known that the assessment of viral load can be highly variable due to both natural biologic variation and assay variation (Brambilla et al., 1999). Since viral load is an important predictor of infectivity, it is possible that our estimates of its association with transmission risk, as

well as our estimates of other covariate parameters, are biased by measurement error. In partner studies we typically have only a single viral load measurement at any given time point, and so we would need to use a different method of adjustment than was used for misreported acts. It may be worth exploring whether regression calibration methods or the SIMEX method could be used in conjunction with our proposed model to adjust for measurement error associated with viral-load or other covariates.

One area of concern with our proposed model is that if the assumed measurement error model for sex acts is incorrect, covariate parameters estimates may be biased. Future work should include a thorough examination of conditions under which the model lacks robustness to mis-specification. Also, it may be valuable to explore whether non-parametric measurement-error models are appropriate, such as those proposed by Carroll et al. (2006).

BIBLIOGRAPHY

Jared M. Baeten, Barbra A. Richardson, Ludo Lavreys, Joel P. Rakwar, Kishorchandra Mandaliya, Job J. Bwayo, and Joan K. Kreiss. Female-to-male infectivity of HIV-1 among circumcised and uncircumcised kenyan men. *Journal of Infectious Disease*, 191: 546–53, 2005.

Jared M. Baeten, Deborah Donnell, Patrick Ndase, Nelly R. Mugo, James D. Campbell, Jonathan Wangisi, Jordan W. Tappero, Elizabeth A. Bukusi, Craig R. Cohen, Elly Katabira, Allan Ronald, Elioda Tumwesigye, Edwin Were, Kenneth H. Fife, James Kiarie, Carey Farquhar, Grace John-Stewart, Aloysious Kakia, Josephine Odoyo, Akasiima Mucunguzi, Edith Nakku-Joloba, Rogers Twesigye, Kenneth Ngure, Harrison Apaka, Cosmas Tamooh, Fridah Gabona, Andrew Mujugira, Dana Panteleeff, Katherine K. Thomas, Lara Kidoguchi, Meighan Krows, Jennifer Revall, Susan Morrison, Harald Haugen, Mira Emmanuel-Ogier, Lisa Ondrejcek, Robert W. Coombs, Lisa Frenkel, Craig Hendrix, Namandj N. Bumpus, David Bangsberg, Jessica E. Haberer, Wendy S. Stevens, Jairam R. Lingappa, and Connie Celum. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine*, 367:399–410, 2012.

BA Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.

Marie Claude Boily, Rebecca F Baggaley, Lei Wang, Benoit Masse, Richard G. White, Richard J. Hayes, and Michel Alary. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *Lancet Infectious Disease*, 9:118–129, 2009.

Donald Brambilla, Patricia Reichelderfer, James Bremer, David Shapiro, Ronald Hershow, David Katzenstein, Scott Hammer, Brooks Jackson, and Ann Collier. The contribution

of assay variation and biological variation to the total variability of plasma HIV-1 RNA measurements. *AIDS*, 1999.

Hermann Brenner. Use and limitations of dual measurements in correcting for nondifferential exposure misclassification. *Epidemiology*, 3(3):216–222, 1992.

Ron Brookmeyer. Analysis of multistage pooling stidies of biological specimens for estimating disease incidence and prevalance. *Biometrics*, 1999.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models; A Modern Perspective*. Chapman and Hall/CRC, 2006.

Connie Celum, Anna Wald, Jairam R Lingappa, Amalia S Margaret, Richard S Wang, Neily Mugo, Andrew Mujugira, Jared M Baeten, James I Mullins, James P Hughes, Elizabeth Bukusi, Craig R Cohen, Elly Katabira, Allan Ronald, James Kiarie, Carey Farquhar, Grace John Stewart, Joseph Makhema, Myron Essex, Edwin Were, Kenneth Fife, Guy de Bruyn, Glenda E Gray, James A McIntyre, Rachel Manongi, Saidi Kapiga, David Coetzee, Susan Allen, Mubiana Inambao, Kayitesi Kayitenkore, Etienne Karita, William Kanweka, Sinead Delany, Helen Rees, Bellington Vwalika, Wendy Stevens, Mary Campell, Katherine K Thomas, Robert W Coombs, Rhoda Morrow, William LH Whittington, Juliana McElrath, Linda Barnes, Renee Ridzon, and Lawrence Corey. Acyclovir and transmission of HIV-1 from persons infected with HIV-1 and HSV-2. *New England Journal of Medicine*, 362(5):427–439, 2010.

Alexander L. Clark and Paul Wallin. The accuracy of husbands' and wives' reports of the frequency of marital coitus. *Population Studies*, 18(2):165–173, 1964.

J Cook and A Stefanski. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1995.

Ingvild Dalen, John P Bounaccorsi, Joseph A Sexton, Petter Laake, and Magne Thoresen. Correction for missclassification of categorized exposure in binary regression using replication data. *Statistics in Medicine*, 28(27):3386–3410, 2009.

AP Dempster, NM Laird, and DB Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 1977.

Viswanath Devanayaran and Leonard A Stefanski. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*, 59:219–225, 2002.

Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 1943.

CC Draper, A Voller, and RG Carpenter. The epidemiologic interpretation of serologic data in malaria. *The American Journal of Tropical Medicine and Hygiene*, 21(5):696–703, 1972.

MJ Evans, Z Gilula, and I Guttman. Latent class analysis of 2-way contingency-tables by bayesian methods. *Biometrika*, 76(3):557–563, 1989.

Kevin A. Fenton, Anne M. Johnson, Sally McManus, and Bob Erens. Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections*, 77:84–92, 2001.

Wayne A. Fuller. *Measurement Error Models*. John Wiley and Sons Inc., 1987.

Richard T Gray, Alexander Hoare, Garrett P Prestage, Basil Donovan, John M Kaldor, and David P Wilson. Frequent testing of highly sexually active gay men is required to control syphilis. *Sexually Transmitted Disease*, 37(5):298–305, 2010.

Ronald H Gray, Maria J Wawer, Ron Brookmeyer, Nelson K Sewankambo, David Serwadda, Fred Wabwire-Mangen, Tom Lutalo, Xianbin Li, Thomas vanCott, Thomas C Quinn, , and the Rakai ProjectTeam. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in rakai, uganda. *Lancet*, 357:1149–1153, 2001.

Paul Gustafson. *Measurement Error and Missclassification in Statistics and Epidemiology; Impacts and Bayesian Adjustments*. Chapman and Hall/CRC, 2004.

Brett Hanscom, Susanne May, and James P Hughes. Efficiently identifying failures using quantitative tests, matrix-pooling and the EM-algorithm. *BePress, UW Biostatistics Working Paper Series*, 2014a.

Brett S Hanscom, James P Hughes, and Susanne May. Avoiding measurement-error bias when estimating per coital act infectivity in discordant partner studies. *Under Review*, 2014b.

K. K. Holmes, D. W. Johnson, and H. J. Trostle. An estimate of the risk of men acquiring gonorrhea by sexual contact with infected females. *American Journal of Epidemiology*, 97:170–174, 1970.

Yosef Hotchberg. On the use of double sampling schemes in analyzing categorical data with misclassification errors. *JASA*, 72:914–921, 1977.

James P. Hughes, Jared M. Baeten, Jairam R. Lingappa, Amalia S. Margaret, Anna Wald, Guy de Bruyn, James Kiarie, Mubiana Inambao, William Kilembe, Carey Farquhar, and Connie Celum. Determinants of per-coital-act HIV-1 infectivity among african HIV-1-serodiscordant couples. *Journal of Infectious Diseases*, 205:358–365, 2012.

Nicholas P. Jewell and Stephan C. Shibowski. Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, 46:1133–1150, 1990.

John Kaldor and David Clayton. Latent class analysis in chronic disease epidemiology. *Statistics in Medicine*, 4:327–335, 1985.

Edward H. Kaplan. Modeling HIV infectivity: must sex acts be counted? *Journal of Acquired Immune Deficiency Syndromes*, 3:55–61, 1990.

MR Kauth, JS St Lawrence, and JA Kelly. Reliability of retrospective assessments of sexual HIV risk behavior: a comparison of biweekly, three-month, and twelve-month self-reports. *AIDS Education and Prevention*, 3:207–214, 1991.

James S. Koopman, Ira M. Longini, John A. Jacquez, Carl P. Simon, David G. Ostrow,

William R. Martin, and David M. Woodcock. Assessing risk factors for transmission of infection. *American Journal of Epidemiology*, 133(12):1199–1209, 1991.

Eugene Litvak. Screening for the presenece of a disease by pooling sera samples. *Journal of the American Statistical Association*, 1994.

Susanne May, Anthony Ganst, Richard Haubrich, Constance Benson, and Davey Smith. Pooled nucleic acid testing to identify anteretroviral treatment failure during HIV infevtion. *Journal of Acquired Immune Defficiency Syndrome*, 2010.

Peter McCulllagh and JA Nelder. *Generalized Linear Models, Second Edition.* Chapman and Hall, 1983.

Mary J Morrissey and Donna Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure. *Biometrics*, 55:338–344, 1999.

NIAID. The jordan report - vaccine updates. Technical report, National Institute Allergy and Infectious Diseases, 2012.

Pragna Patel, Craig B. Borkowf, John T. Brooks, Arielle Lasry, Any Lansky, and Jonathan Mermin. Estimating per-act HIV transmission risk: a systematic review. *AIDS*, 28: 1509–1519, 2014.

RM Phadarfod and Aidan Sudbury. The use of a square array scheme in blood testing. *Statistics in Medicine*, 1994.

Prinitha Pillay, Zara Shubber, and Rashida A Ferrand. Outcomes for efavirenz versus nevirapine-containing regimens for treatment of HIV-1 infection: A systematic review and meta-analysis. *PLOS ONE*, 2013.

Richard Platt, Peter A Rice, and William M McCormack. Risk of acquiring gonorrhea and prevalence of abnormal adnexal findings among women recently exposed to gonorrhea. *Journal of the Americal Medical Association*, 250:3205–3209, 1983.

Kimberly A Powers, Charles Poole, Audrey E Pettifor, and Myron S. Cohen. Rethinking the heterosexual infectivity of HIV-1: a systematic review and meta-analysis. *Lancet Infectious Disease*, 8:553–563, 2008.

S Richardson and WR Gilks. Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, 12:1703–1722, 1993a.

S Richardson and WR Gilks. A bayesian-approach to measurement error problems in epidemiology using conditional-independence models. *American Journal of Epidemiology*, 138:430–442, 1993b.

David Rindskopf and Wallace Rindskopf. The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5:21–27, 1986.

Michael Rosenblum, Steven G Deeks, Mark van der Laan, and David R Bangsberg. The risk of virologic failure decreases with duration of HIV suppression, at greater than 50 percent adherence to antiretroviral therapy. *PLOS ONE*, 2009.

B Rosner, D Spiegelman, and WC Willett. Correction of logistic-regression relative risk estimates and confidence-intervals for random within-person measurement error. *American Journal of Epidemiology*, 136:1400–1413, 1992.

Glen A. Satten, Timothy D. Mastro, and Ira M. Longini. Modelling the female-to-male per-act HIV transmission probability in an emerging epidemic in asia. *Statistics in Medicine*, 13:2097–2106, 1994.

DW Schafer. Likelihood analysis and flexible structural modeling for measurement error model regression. *J. Statist. Computn Simuln*, 72:33–45, 2002.

Stephan C. Shibowski and Nicholas P. Jewell. Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, 87(418):360–372, 1992.

DM Smith, Susanne May, Josue Perez-Santiago, MC Strain, CC Ignacio, RH Haubrich,

DD Richmana, CA Benson, and SJ Little. The use of pooled viral load testing to identify antiretroviral treatment failure. *AIDS*, 2009.

Manueko Sobel and RM Elashoff. Group testing with a new goal, estimation. *Biometrika*, 1975.

Andrew Sterrett. On the detection of defective members of large populations. *Annals of Mathematical Statistics*, 1957.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 1987.

Duncan C Thomas. Multistage sampling for latent variable models. *Lifetime Data Analysis*, 13:565–581, 2007.

KH Thompson. Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, 1962.

UNAIDS. The gap report. Technical report, Joint United Nations Programme on HIV/AIDS, 2014.

GH van Zyl, W Preiser, S Potschka, AT Lundershaisen, R Haubrich, and D Smith. Pooling strategies to reduce the cost of HIV-1 RNA load monitoring in a resource-limited setting. *HIV/AIDS*, 2011.

N Wang, X Lin, R Gutierrez, and Raymond J Carroll. SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93(418):249–261, 1999.

Zathleen A Wannemuehler. A unified model for covariate measurement error adjustment in an occupational health study while accounting for non-detectable exposures. *Journal of the Royal Statistical Scociety Series C (Applied Statistics)*, 54:259–271, 2005.

Maria J. Wawer, Ronald H. Gray, Nelson K. Sewankambo, David Serwadda, Xianbin Li, Oliver Laeyendecker, Noah Kiwanuka, Godfrey Kigozi, Mohammed Kiddugavu, Thomas

Lutalo, Fred Nalugoda, Fred Wabwire-Mangen, Mary P. Meehan, and Thomas C. Quinn3. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in rakai, uganda. *The Journal of Infectious Diseases*, 191:1403–1409, 2005.

Ian White, Chris Frost, and Shoji Tokunaga. Correcting for measurement error in binary and continuous variables using replicates. *Statistics in Medicine*, 20:3441–3457, 2001.

WHO. Access to antiretroviral drugs in low- and middle-income countries. Technical report, World Health Organization, 2014a.

WHO. Hiv/aids fact sheet no. 360. Technical report, World Health Organization, 2014b.

M Xie, K Tatsouka, J Sacks, and SS Young. Group testing with blockers and synergism. *JASA*, 2001.

Yang Yang, Peter Gilbert, Ira M. Longini, and M. Elizabeth Halloran. A bayesian framework for estimating vaccine efficacy per infectious contact. *The Annals of Applied Statistics*, 2 (4):1409–1431, 2008.

VITA

Brett was born in Salt Lake City, UT in 1970. He graduated from Middlebury College in 1992 with a Bachelors degree in Mathematics, and then earned a Masters degree in Statistics and Econometrics from the University of Utah in 1995. He worked for the Utah Office of Energy and Resource Planning for two years and Dartmouth Medical School for ten years. In 2007 he moved to Seattle to pursue a PhD in Biostatistics.