11-2012

GROUP TESTING REGRESSION MODELS

Boan Zhang

University of Nebraska-Lincoln, boan.zhang@huskers.unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/statisticsdiss

Part of the Biostatistics Commons, Categorical Data Analysis Commons, and the Statistical Methodology Commons

http://digitalcommons.unl.edu/statisticsdiss/9

This Article is brought to you for free and open access by the Statistics, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Dissertations and Theses in Statistics by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
GROUP TESTING REGRESSION MODELS

by

Boan Zhang

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor Christopher R. Bilder

Lincoln, Nebraska

November, 2012
Group Testing Regression Models

Boan Zhang, Ph.D.
University of Nebraska, 2012

Advisor: Christopher R. Bilder

Group testing, where groups of individual specimens are composited to test for the presence or absence of a disease (or some other binary characteristic), is a procedure commonly used to reduce the costs of screening a large number of individuals. Statistical research in group testing has traditionally focused on a homogeneous population, where individuals are assumed to have the same probability of having a disease. However, individuals often have different risks of positivity, so recent research has examined regression models that allow for heterogeneity among individuals within the population. This dissertation focuses on two problems involving group testing regression models.

For the first problem, we examine group testing regression models when identification of the positive and negative statuses for individuals is performed. The identification aspect leads to additional tests, known as “retests,” beyond those performed for initial groups of individuals. We show how regression models can be fit in this setting while also incorporating the extra information from these retests. Through Monte Carlo simulations, we present evidence that significant gains in efficiency occur by incorporating retesting information. Furthermore, we demonstrate that some group testing protocols can actually lead to more efficient estimates than individual testing when diagnostic tests are imperfect. Finally, we
show that halving and matrix testing protocols are the most efficient to use in application.

For the second problem, we consider situations when individuals are tested in groups for multiple diseases simultaneously. This problem is important because assays frequently screen for more than one disease at a time. When these assays are used in a group testing setting, the individual positive/negative statuses consist of unobserved, correlated random variables. To estimate models in this setting, we develop an expectation-solution based algorithm that provides consistent parameter estimates and natural large-sample inference procedures.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Christopher R. Bilder, from the bottom of my heart for his support, patience, and devotion, and for putting so much effort and time into guiding me. I would also like to thank my committee members, Dr. Kent Eskridge, Dr. Dong Wang, Dr. Joshua Tebbs, and Dr. Stephen Hartke for their tremendous help, guidance, and support throughout my graduate student years. Finally, I want to express my deepest gratitude and love for my wonderful parents for their love, care, and support throughout my student years.
# Table of Contents

Chapter 1: Introduction .................................................................................................................. 1  
1.1 Background ............................................................................................................................... 1  
1.2 Group Testing Regression Models ............................................................................................ 2  
1.3 Group Testing Protocols used for Identification ..................................................................... 5  
1.3.1 Dorfman .............................................................................................................................. 6  
1.3.2 Halving ............................................................................................................................... 6  
1.3.3 Array testing ....................................................................................................................... 7  
1.4 Motivation and Objectives ....................................................................................................... 7  
1.5 Organization of the Dissertation ............................................................................................. 9  

Chapter 2: Paper #1 - Group Testing Regression Model Estimation when Case Identification is a Goal ......................................................................................................................... 12  
2.1 Introduction ............................................................................................................................. 13  
2.2 Estimation of Group Testing Regression Models ..................................................................... 15  
2.2.1 Initial Group Tests from Non-Overlapping Groups ......................................................... 17  
2.2.2 Dorfman ............................................................................................................................. 18  
2.2.3 Halving ............................................................................................................................... 19  
2.2.4 Array Testing .................................................................................................................... 20  
2.3 Simulation Study .................................................................................................................... 23  
2.3.1 Estimator Accuracy and Variance Estimation ................................................................. 24  
2.3.2 Improvements in Variance Estimation from Including Retests ...................................... 25  
2.3.3 Average Number of Tests per Unit of Information ......................................................... 27  
2.3.4 Additional Simulations ...................................................................................................... 28  
2.4 Infertility Prevention Project ................................................................................................. 28  
2.5 Discussion ............................................................................................................................... 30
List of Tables

Table 2.1. The numerator and denominator for $\omega_{ik}$ for the halving protocol in Section 2.2.3. .................................................................35

Table 2.2. Parameter estimates and their standard errors based on 1000 (300 for array testing) simulated data sets in Section 2.3.2 .........................36

Table 2.3. Parameter estimates and estimated standard errors for the chlamydia screening data in Section 2.4 ..........................................................37

Table 3.1. Simulation results from using the ES algorithm to estimate the model in Section 3.4 .............................................................................64

Table 3.2. Relative efficiency of the variance estimates for the model in Section 3.4 .........................................................................................65

Table 3.3. Parameter estimates and estimated standard errors for the NPHL data. .........................................................................................66

Table 4.1. GEE-group simulation results corresponding to the model in Section 3.4 .........................................................................................82

Table I.1. Comparison of $\alpha$ estimates averaged over 5 simulated data sets......118

Table J.1. Relative efficiency of the variance estimates for the model Section 3.4 when $\beta_{21}$ and $\beta_{22}$ are estimated separately .................................120

Table J.2. Relative efficiency of the variance estimates for additional simulations in Appendix J ..............................................................................121

Table K.1. Parameter and standard error estimates for the parsimonious model described in Section 3.5 ...............................................................122
List of Figures

Figure 2.1. Relative efficiencies calculated by Equation (6) based on 500 simulated data sets in Section 2.3.2 .................................38

Figure 2.2. Averaged $Var(\hat{\beta}_1)$ for 500 simulated data sets in Section 2.3.2 ......39

Figure 2.3. Average number of tests per unit of information calculated by Equation (7) based on 500 simulated data sets in Section 2.3.3. .................40

Figure 4.1. True correlation between multiple disease responses plotted over the groups for the first set of simulations in Section 4.4...............................83

Figure D.1. Histogram of the true individual probabilities for one simulated data set in Section 2.3......................................................................................106

Figure F.1. Relative efficiencies calculated by Equation (6) based on the additional simulations in Appendix F.................................................................106

Figure F.2. Averaged $Var(\hat{\beta}_1)$ for the simulations in Appendix F.................107

Figure F.3. Average number of tests per unit of information calculated by Equation (7) based on the simulations in Appendix F .........................108

Figure F.4. A histogram of the true individual probabilities for one simulated data set in Appendix F......................................................................................108

Figure J.1. Histograms of the true individual probabilities for one Section 3.4 simulated data set ......................................................................................119
Chapter 1: Introduction

1.1 Background

When performing individual testing for an infectious disease, a specimen (e.g., blood, urine) is obtained from each subject and tested to determine the positive or negative disease status of the subject. When there are a large number of specimens and/or testing costs are high, this can result in excessive time and expenditures to complete the screening process. In these situations, group testing, where individual specimens are composited into pools for testing, has become standard practice. If a pool tests negative, all individuals within it are diagnosed as being negative. If a pool tests positive, there is at least one positive individual within it; further retesting of those individuals can be completed to determine individual diagnoses. As long as the overall prevalence of the disease is low and appropriate group sizes are chosen, group testing can considerably reduce the number of tests and associated costs when compared to individual testing.

The use of group testing in screening people for low-prevalence diseases has a long history dating back to detecting syphilis among World War II soldiers (Dorfman 1943). Subsequently, group testing has been successfully adopted in many infectious disease applications, including blood donation screening by the American Red Cross (American Red Cross 2012), opportunistic chlamydia and gonorrhea testing in medical clinics (Gaydos 2005), and Bovine Viral Diarrhea virus detection for the cattle industry (Munoz-Zanzi et al. 2006). Group testing has also proven to be beneficial in areas outside of infectious disease prevalence estimation and detection, including drug discovery experiments (Remlinger et al.
2006), plant pathology (Tebbs and Bilder 2004), genotyping (Chi et al. 2009), and food contamination testing (Fahey, Ourisson, and Degnan 2006).

1.2 Group Testing Regression Models

Group testing is generally used for two purposes: case identification and prevalence estimation. In the context of infectious diseases, the goal of case identification is to identify all individuals having a disease. In contrast, the goal of prevalence estimation is to estimate the prevalence of a disease in a population. The focus of this dissertation is prevalence estimation, and most research in this area has examined estimating an overall prevalence in a homogeneous population (e.g., Swallow 1985; Biggerstaff 2008; Hepworth and Watson 2009). However, populations are frequently heterogeneous, where covariates, such as gender, behavior, education level, ..., may influence individual disease status. For this reason, it is important to estimate covariate-specific probabilities of positivity. In the remainder of this section, we review the two seminal works in this area – one by Vansteelandt et al. (2000) and one by Xie (2001) – that have been proposed to estimate these probabilities through group testing regression models.

Define $\tilde{Y}_{ik} = 1$ if the $i^{th}$ individual in the $k^{th}$ group is positive and $\tilde{Y}_{ik} = 0$ otherwise, for $i = 1, ..., I_k$, $k = 1, ..., K$. We assume that each individual is assigned to exactly one group and that $\tilde{Y}_{ik}$ are independent Bernoulli random variables. Denote the observed group response for the $k^{th}$ group as $Z_k$. If group tests are perfectly accurate, a group tests positive $(Z_k = 1)$ if and only if $\sum_{i=1}^{I_k} \tilde{Y}_{ik} > 0$ and a group tests negative $(Z_k = 0)$ if and only if $\sum_{i=1}^{I_k} \tilde{Y}_{ik} = 0$. The group responses $Z_k$ are also independent Bernoulli random variables with mean $\theta_k \equiv P(Z_k = 1)$. 
Because assays are usually subject to testing errors, the sensitivity and specificity of a test are defined as \( \eta = P(Z_k = 1 | \tilde{Z}_k = 1) \) and \( \delta = P(Z_k = 0 | \tilde{Z}_k = 0) \), respectively, where \( \tilde{Z}_k \) denotes the true response of group \( k \). Vansteelandt et al. (2000) and Xie (2001) both assumed that the sensitivity and specificity are known and do not depend on pool size; these assumptions are reasonable with properly calibrated modern diagnostic assays. Using the total probability theorem, one can express \( \theta_k \) in terms of the true individual probabilities \( \tilde{p}_{ik} \) as

\[
\theta_k = P(Z_k = 1 | \tilde{Z}_k = 1)P(\tilde{Z}_k = 1) + P(Z_k = 1 | \tilde{Z}_k = 0)P(\tilde{Z}_k = 0) \\
= \eta + (1 - \delta - \eta) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ik}).
\]

To incorporate risk factors that may influence an individual’s response, our goal is to estimate \( \tilde{p}_{ik} \) as a function of the covariates for each individual. The model of interest is

\[
f(\tilde{p}_{ik}) = \beta_0 + \beta_1 x_{i1k} + \ldots + \beta_{p-1} x_{p1ik},
\]

where \( f(\cdot) \) is a known, monotonic, differentiable link function and \( x_{i1k}, \ldots, x_{p1ik} \) are the \( p - 1 \) covariates for the \( i^{th} \) individual in the \( k^{th} \) group.

As mentioned earlier, Vansteelandt et al. (2000) and Xie (2001) present two different ways to find the maximum likelihood estimates for \( \beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})' \). Using Equation (1), Vansteelandt et al. (2000) writes the likelihood function in terms of observed group responses \( z_k' \):

\[
L_k = \prod_{k=1}^{K} \theta_k^{z_k} (1 - \theta_k)^{1-z_k} \\
= \prod_{k=1}^{K} \left[ \eta + (1 - \eta - \delta) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ik}) \right]^{z_k} \left[ 1 - \eta - (1 - \eta - \delta) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ik}) \right]^{1-z_k}
\]
for the model in (2). Maximizing $L_1$ directly results in the parameter estimates $\hat{\beta}$. The covariance matrix for $\hat{\beta}$ can be obtained from the inverse of the observed Fisher information matrix.

Alternatively, Xie (2001) expresses the likelihood function in terms of the unobserved individual responses $\tilde{y}_{ik} : L_2 = \prod_{k=1}^{K} \prod_{i=1}^{l_k} \tilde{p}_{ik}^{\tilde{y}_{ik}} (1 - \tilde{p}_{ik})^{1-\tilde{y}_{ik}}$, and proposed the use of an EM algorithm in maximizing the likelihood function. The algorithm works by replacing the unobserved outcomes $\tilde{y}_{ik}$ in $\log(L_2)$ by $\omega_{ik} = E(\tilde{Y}_{ik} | I)$ where $I$ denotes all information obtained by group tests and retests under a particular group testing protocol. The expectation and maximization steps of the EM algorithm are alternated between in an iterative manner until convergence is reached. Formally, the following EM algorithm can be used to obtain the maximum likelihood estimate of $\beta$, denoted by $\hat{\beta}$:

1) Select a starting point $\beta^{(0)}$ of $\beta$.

2) E-step: For a given $\beta^{(b)}$, $b = 0, 1, 2, \ldots$, calculate $\omega^{(b)}_{ik} = E(\tilde{Y}_{ik} | I, \beta^{(b)})$, for $i = 1, \ldots, I_k$ and $k = 1, \ldots, K$.

3) M-step: Maximize the following function

$$E[\log(L_2) | z_1, \ldots, z_K, \beta^{(b)}] = \sum_{k=1}^{K} \sum_{i=1}^{l_k} \omega^{(b)}_{ik} \log(\tilde{p}_{ik}) + (1 - \omega^{(b)}_{ik}) \log(1 - \tilde{p}_{ik})$$

for $\beta$ to update the parameter estimates at the $(b + 1)^{th}$ iteration.

4) Repeat Steps 2 and 3 until $\|\beta^{(b+1)} - \beta^{(b)}\|$ is very small; denote the final solution by $\hat{\beta}$.

To estimate the covariance matrix of $\hat{\beta}$, Louis’s (1982) method can be applied to obtain the Hessian matrix:
\[ H_n = \sum_{j=1}^{N} \left( \omega_j - \frac{\partial^2}{\partial \beta \partial \beta'} \left[ \log \left( \frac{\tilde{p}_j}{1 - \tilde{p}_j} \right) \right] + \frac{\partial^2}{\partial \beta \partial \beta'} \left[ \log(1 - \tilde{p}_j) \right] \right) - \]

\[ \sum_{j=1}^{N} \sum_{j' \neq j} (\omega_{j'} - \omega_j) \left( \frac{\partial}{\partial \beta} \left[ \log \left( \frac{\tilde{p}_j}{1 - \tilde{p}_j} \right) \right] \left( \frac{\partial}{\partial \beta} \left[ \log \left( \frac{\tilde{p}_{j'}}{1 - \tilde{p}_{j'}} \right) \right] \right) \right)' , \]

where we have re-indexed the subjects to be \( j = 1, \ldots, N, N = \sum_{k=1}^{K} I_k \), \( \omega_j = E(\tilde{Y}_j | \mathcal{I}) \), and \( \omega_{j'} = E(\tilde{Y}_j \tilde{Y}_{j'} | \mathcal{I}) \). The inverse of \( \hat{H}_n \), which is \( H_n \) evaluated at \( \hat{\beta} \), is used as the covariance matrix estimate.

Xie (2001) remarks that closed form expressions for \( \omega_{ik} \) are not possible for some group testing protocols (algorithm used for the initial testing and subsequent retesting). In these cases, a Gibbs sampling approach can be employed to estimate them; we will illustrate such technique in Section 2.2.4. Because \( \mathcal{I} \) can include information from any group or individual tests, Xie's (2001) method is very flexible and can deal with a wide range of complex group testing protocols.

Several very recent papers have expanded on the work of these two seminal papers in this area. Specifically, Bilder and Tebbs (2009) provide a thorough comparison of individual and group testing regression model estimates. Chen et al. (2009) examine mixed-effects models, and Delaigle and Meister (2011) and Delaigle and Hall (2012) discuss a nonparametric modeling approach.

### 1.3 Group Testing Protocols Used for Identification

In this sub-section, we introduce group testing protocols that are commonly used for case identification. In each of the protocols described next, there may be multiple responses involving each individual due to retests, which makes a direct evaluation of a likelihood function difficult. In later chapters, we will investigate
how to incorporate the additional retest information from these protocols into estimating the model given in (2).

1.3.1 Dorfman

Although group testing had been used earlier (see Hughes-Oliver (2006) for a review), Dorfman (1943) is largely regarded as the seminal paper in the area. Dorfman proposed screening pooled blood samples of US Army soldiers for syphilis, followed by retesting all soldiers individually within positive pools. Individuals within negative testing groups were declared negative. Due to its simplicity, Dorfman’s protocol is the most widely adopted protocol for case identification, and its applications include screening blood donations (Stramer et al. 2004), chlamydia screening (Mund et al. 2008), and potato virus detection (Liu et al. 2011).

1.3.2 Halving

The halving protocol is an alternative to Dorfman’s protocol, where positive testing groups are successively split into two equal sized subgroups. If a subgroup tests negative, no further splitting is needed and its members are declared negative; if a subgroup tests positive, it is further split and tested until all subgroups test negative or until individual testing occurs. For example, the first step of a 3-stage halving protocol with a group of size $I = 8$ is to test the whole group. If the group tests positive, it is split into two subgroups of size 4. If either subgroup tests positive, all individuals within a positive subgroup are tested. If a 4-stage halving protocol is used instead, we would have one more round of splitting the subgroups before individual testing. For simplicity, we only consider

1.3.3 Array testing

The array testing protocol, first proposed by Phatarfod and Sudbury (1994), assigns individuals to overlapping groups arranged into a two-dimensional array structure. Specimens are pooled within each row and within each column for testing. Intersections of positive testing rows and columns indicate where positive individuals may exist. When more than one row and more than one column test positive, ambiguities arise on which of these individuals at the intersections led to the positive row and column test results. We may also have one or more rows testing positive and no columns testing positive (or vice versa) when testing errors are present. To clear these ambiguities, additional testing (usually on each individual) can be used to complete the decoding. The array testing protocol has found much success in high throughput screening applications, such as infectious disease testing (Tilghman et al. 2011), DNA screening (Berger et al. 2000), and systems biology (Thierry-Mieg 2006).

1.4 Motivation and Objectives

The Centers for Disease Control and Prevention and the Office of Population affairs support the Infertility Prevention Program (IPP) in order to reduce the prevalence of chlamydia and gonorrhea in the United States, while also to better understand factors affecting prevalence (Centers for Disease Control and
Prevention, 2012). Each state participates in the IPP. In Nebraska, health care clinics across the state obtain urine and swab specimens to test for the diseases. These specimens are sent to the Nebraska Public Health Laboratory (NPHL) for testing where in total approximately 25,000 tests are performed yearly for these diseases. Along with the specimens, each individual screened contributes a set of information, such as age, gender, symptoms, and past history of risky behavior. Clinical observations are made as well on each individual, including cervical friability, pelvic inflammatory disease, cervicitis, and urethritis statuses.

All current testing at the NPHL is performed individually on each specimen; i.e., group testing is not used. Due to the large number of specimens screened annually and the high cost associated with these tests (approximately $11 for a swab test and $16 for a urine test), group testing could be very efficient and beneficial if employed by the NPHL. In particular, in order to understand how certain risk factors influence the disease statuses, one can fit a group testing regression model that estimates an individual’s probability of having chlamydia or gonorrhea at a largely reduced cost, as compared to testing specimens individually. In later chapters, we will detail how group testing could be used by the NPHL and the benefits associated with its use.

In many applications, prevalence estimation and case identification are simultaneous goals. For example, the goals of the IPP involve not only the identification of positive individuals, but also to evaluate risk factors closely related to infection. Even in public health studies where prevalence estimation is the primary goal and only the initial group tests are needed, retests on individuals are frequently performed for ethical reasons. Although Xie (2001) proposed the general EM algorithm framework for group testing regression
problems, details on how to implement his proposal for specific group testing protocols were not given for any of the commonly used protocols described in Section 1.3. Furthermore, it is unknown which protocol results in the more efficient estimators. In Chapter 2, we examine how the general EM algorithm of Xie (2001) can be applied to these three group testing protocols introduced in Section 1.3 and develop recommendations on their use.

In practice, there are many cases where testing is done not only for one disease, but for multiple diseases at the same time. For example, one assay is used at the NPHL to test for chlamydia and gonorrhea simultaneously. Also, the American Red Cross screens blood donations for HIV, hepatitis B, hepatitis C, and West Nile Virus through using group testing (American Red Cross 2012; Dodd et al. 2002; Stramer et al. 2004). With respect to group testing, Hughes-Oliver and Rosenberger (2000) is the lone paper that addresses the multiple-disease problem, and they only examined the homogeneous population situation. The purpose of Chapter 3 then is to take advantage of covariate information to model individual statuses of multiple diseases simultaneously in a group testing setting. In general, we are proposing a regression model for unobserved correlated binary responses.

1.5 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 is a paper under review by Biometrical Journal. The paper shows how regression models can be fit to group testing data from three commonly used group testing protocols: Dorfman, halving, and array testing, as described in Section 1.3. Simulation evidence is presented to show significant efficiency gains from
incorporating retests into the estimation process, as compared to using the initial group test results alone. We also discover that group testing with retests can result in more efficient estimators than individual testing when testing error is present. Thus, not only will group testing lead to a smaller number of tests, but more information can be gained by using group testing. Finally, we investigate which group testing protocol leads to the most efficient estimators overall.

Chapter 3 contains almost all of a paper that is under review at *Statistics in Medicine* (an additional example, set of simulations, and parts of the paper’s discussion section were completed by my advisor, so they are omitted from the dissertation). In this paper, we propose the first regression techniques for multiple-disease group testing data. We develop an expectation-solution based algorithm that takes into account the correlation structure of unobserved individual disease statuses. Simulation studies show the consistency of our estimators as well as efficiency gains in parameter estimates when compared to single-disease group testing models.

Chapter 4 includes additional research completed that did not fit into Chapters 2 and 3. We show how to generalize the model-fitting procedure of Chapter 3 to incorporate individual retesting information. We also present alternative approaches to the methods proposed in Chapters 2 and 3. To conclude, we present a discussion of future directions for research involving group testing regression models.

Both Chapters 2 and 3 contain the references as given in their corresponding papers. We also include all references cited throughout the dissertation in a separate references section toward the end of the dissertation. All appendices are
located at the end of the dissertation. These include appendices that were “web appendices” for the paper submissions.
Chapter 2: Paper #1 - Group Testing
Regression Model Estimation when Case Identification is a Goal

Abstract

Group testing is frequently used to reduce the costs of screening a large number of individuals for infectious diseases or other binary characteristics in small prevalence situations. In many applications, the goals include both identifying individuals as positive or negative and estimating the probability of positivity. The identification aspect leads to additional tests being performed, known as “retests,” beyond those performed for initial groups of individuals. In this paper, we investigate how regression models can be fit to estimate the probability of positivity while also incorporating the extra information from these retests. We present simulation evidence showing that significant gains in efficiency occur by incorporating retesting information. Furthermore, we demonstrate that some group testing protocols can actually lead to more efficient estimates than individual testing when diagnostic tests are imperfect. Finally, we examine which protocols are the most efficient to use in application. Our methods are illustrated using chlamydia screening data from the Infertility Prevention Project.

Key words: Binary response; Generalized linear model; EM algorithm; Latent response; Pooled testing; Prevalence estimation.
2.1 Introduction

Pooling specimens to screen a population for infectious diseases has a long history dating back to Dorfman’s (1943) proposal to screen American soldiers for syphilis during World War II. Today, testing individuals in pools through group testing (also known as “pooled testing”) has been successfully adopted in many additional areas, including entomology (Gu et al. 2004), veterinary medicine (Muñoz-Zanzi et al. 2000), DNA screening (Berger et al. 2000), and drug discovery (Kainkaryam and Woolf 2009). When compared to testing specimens individually, group testing can provide considerable savings in time and costs when the overall prevalence of the disease (or some other binary characteristic of interest) is low. This makes the use of group testing particularly desirable in applications where there are limitations in resources.

Group testing is generally used for two purposes: case identification and prevalence estimation. The goal of case identification is to identify all individuals as being positive or negative. Individual specimens are initially pooled into groups, and these groups are tested. Individuals within positive testing groups are then retested in some prior specified way to distinguish positive individuals from those that are negative. The goal of prevalence estimation is to estimate the prevalence of positivity in a population. Retesting is not needed in this case because initial group test responses alone can be used to estimate the prevalence. However, when prevalence estimation and case identification are simultaneous goals, the additional retesting information can be used for estimation as well. Intuitively, one would expect statistical benefits (e.g., in terms of efficiency) from
including retest outcomes as part of the estimation process. Our paper examines how to include retests while also quantifying the benefits from their inclusion.

The majority of group testing estimation research has focused on inference for an overall prevalence $p$ using only the results from the initial group tests (e.g., Swallow 1985; Biggerstaff 2008; Hepworth and Watson 2009). A few papers, such as Sobel and Elashoff (1975) and Chen and Swallow (1990), discuss including retests to estimate $p$, but under the restriction of perfect testing and without positive case identification. More recently, estimation research has focused on regression modeling to obtain an estimate of individual positivity, given a set of risk factors. The seminal papers in this area, Vansteelandt et al. (2000) and Xie (2001), both propose likelihood-based estimation and inference using binary regression models, but their approaches differ. Vansteelandt et al. (2000) use a likelihood function written in terms of the initial group responses, and standard techniques for generalized linear models are used to find the parameter estimates that maximize this function. Xie (2001) uses a likelihood function written in terms of the true latent individual statuses and then employs the EM algorithm to maximize the likelihood function. The main advantage of Xie’s approach is that it allows for the inclusion of retests.

Given the large number of ways to retest individuals within positive groups (see Hughes-Oliver (2006) for a review), it is important to determine if there are benefits from including retest outcomes when estimating a group testing regression model. The purpose of our paper is to determine if benefits truly exist, and, in particular, determine which group testing protocol (algorithm used for the initial testing and subsequent retesting) is the most efficient. This is especially important because group testing is typically applied in settings where cost and
time considerations are a primary concern. Ideally, one would want to apply a protocol that results in the fewest number of tests while also producing the most efficient regression estimates. Also, model estimation plays a significant role in the application of informative retesting procedures for case identification (e.g., see Bilder et al. (2010) and Black et al. (2012)). These identification procedures rely on group testing regression models to identify which individuals are most likely to be positive, so having the best possible estimates is crucial.

The order of our paper is as follows. Section 2.2 reviews three commonly used group testing protocols. Note that each of these protocols are not specifically examined in Xie (2001), so this is the first time that the EM algorithm details have been formally presented for them. In Section 2.3, we use simulation to investigate the benefits from including retests and determine which protocol is the most efficient. This section also shows that group testing can actually be more efficient than individual testing when estimating regression parameters. In Section 2.4, we apply these protocols to chlamydia screening data from the Infertility Prevention Project, where both identification and prevalence estimation are important. Finally, Section 2.5 summarizes our findings and discusses extensions to this research.

2.2 Estimation of Group Testing Regression Models

Define $\tilde{Y}_{ik} = 1$ if the $i^{th}$ individual in the $k^{th}$ initial group is truly positive and $\tilde{Y}_{ik} = 0$ otherwise, for $i = 1, \ldots, I_k$ and $k = 1, \ldots, K$. Our goal is to estimate $E(\tilde{Y}_{ik}) = \tilde{p}_{ik}$, conditional on a set of covariates $x_{1ik}, \ldots, x_{p-1,ik}$, using the regression model

$$f(\tilde{p}_{ik}) = \beta_0 + \beta_1 x_{1ik} + \ldots + \beta_{p-1} x_{p-1,ik},$$

(3)
where $f(\cdot)$ is a known monotonic, differentiable function. The log-likelihood function can be written as

$$
\log[L(\beta)] = \sum_{k=1}^{K} \sum_{i=1}^{L} \hat{y}_{ik} \log(\hat{p}_{ik}) + (1 - \hat{y}_{ik}) \log(1 - \hat{p}_{ik}),
$$

(4)

where $\beta = (\beta_0, \ldots, \beta_{p-1})'$ and we assume that the $\hat{Y}_{ik}$ are independent Bernoulli($\hat{p}_{ik}$) random variables. If the true individual statuses $Y_{ik}$ were observed, likelihood-based estimation for the model would proceed in a straightforward manner.

In group testing applications, the individual statuses $Y_{ik}$ are unknown because only group responses may be observed and because groups and/or individuals may be misclassified due to diagnostic testing error. To fit the model, Xie (2001) proposed the use of an EM algorithm to find the parameter estimates that maximize the likelihood function. The algorithm works by replacing the unobserved outcomes $y_{ik}$ in Equation (4) by $\omega_{ik} = E(\hat{Y}_{ik} | I)$, where $I$ denotes all information obtained by group tests and retests under a particular group testing protocol. The expectation and maximization steps of the EM algorithm are alternated between in an iterative manner until convergence is reached to obtain the maximum likelihood estimate of $\beta$, denoted by $\hat{\beta}$. The estimated covariance matrix of $\hat{\beta}$ is obtained by standard methods; e.g., see Louis (1982) and Xie (2001, p. 1960).

The most difficult aspect of the EM algorithm application is to obtain the conditional expectations $\omega_{ik}$. Xie (2001) provides derivation details only for the protocol outlined in Gastwirth and Hammick (1989), which involves testing individuals in non-overlapping groups and performing one confirmatory test on groups that test positive. While this protocol can be extremely useful for
estimation purposes, it can not be used to identify positive individuals. In this paper, we consider three group testing protocols commonly used in practice for case identification. The following subsections elaborate on how to calculate the conditional expectations $\omega_{ik}$ for each protocol. Given these details, the EM algorithm for fitting Equation (3) becomes straightforward to implement.

### 2.2.1 Initial Group Tests from Non-Overlapping Groups

Initial tests from groups that are non-overlapping (i.e., each individual is within only one group) do provide enough information to estimate Equation (2), although not as efficiently as other case identification protocols to be discussed shortly. We begin by describing how models can be fit under this setting to motivate model fitting when retests are included.

Define $Z_k$ as the response for initial group $k$, where $Z_k = 1$ denotes a positive test result and $Z_k = 0$ denotes a negative test result. Because diagnostic tests are likely subject to error, we define the true status of a group by $\tilde{Z}_k$ where a 1 (0) again denotes a positive (negative) status. The sensitivity and specificity of the group test are given by $\eta = P(Z_k = 1 \mid \tilde{Z}_k = 1)$ and $\delta = P(Z_k = 0 \mid \tilde{Z}_k = 0)$, where we assume these values are known and do not depend on group size. These assumptions are consistent with most research for group testing regression, including Vansteelandt et al. (2000) and Xie (2001). When only the initial group responses are observed, $\omega_{ik}$ is easily found as

$$
\omega_{ik} = \begin{cases} 
P(\tilde{Y}_{ik} = 1 \mid Z_k = 0) = (1 - \eta) \hat{p}_{ik} / (1 - \theta_k), & \text{if } Z_k = 0 \\
P(\tilde{Y}_{ik} = 1 \mid Z_k = 1) = \eta \tilde{p}_{ik} / \theta_k, & \text{if } Z_k = 1, 
\end{cases}
$$

where
\[ \theta_k = P(Z_k = 1 \mid \tilde{Z}_k = 1)P(\tilde{Z}_k = 1) + P(Z_k = 1 \mid \tilde{Z}_k = 0)P(\tilde{Z}_k = 0) = \eta + \varphi \prod_{i=1}^{l} (1 - \tilde{p}_ik) \]

is the probability that group \(k\) tests positive and \(\varphi = 1 - \eta - \delta\).

### 2.2.2 Dorfman

After initially testing individuals in non-overlapping groups, Dorfman (1943) proposed to individually retest all specimens within the positive testing groups. Individuals within negative testing groups are declared negative. Because of its simplicity, Dorfman’s protocol is the most widely adopted protocol for case identification, and its applications include screening blood donations (Stramer et al. 2004), chlamydia testing (Mund et al. 2008), and potato virus detection (Liu et al. 2011).

Because specimens are retested, \(\omega_{ik}\) is no longer the same as given in Equation (5) when a group tests positive. Let \(Y_{ik}\) denote the retest outcome for individual \(i\) in group \(k\) and assume that the same assay for group tests is also used for individual retests (thus, \(\eta\) is the sensitivity and \(\delta\) is the specificity for properly calibrated tests). For observed positive groups \((Z_k = 1)\), we have calculated

\[
\omega_{ik} = P(\tilde{Y}_{ik} = 1 \mid Y_{ik} = y_{ik}, \ldots, Y_{l,k} = y_{l,k}, Z_k = 1) \\
= \frac{\hat{p}_ik \eta P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \prod_{i=1}^{l} \sum_{\tilde{y}_{ik} = 0}^{1} P(Y_{\tilde{i},k} = y_{\tilde{i},k} \mid \tilde{Y}_{\tilde{i},k} = \tilde{y}_{\tilde{i},k}) P(\tilde{Y}_{\tilde{i},k} = \tilde{y}_{\tilde{i},k})}{\left\{ \varphi \prod_{i=1}^{l} P(Y_{\tilde{i},k} = y_{\tilde{i},k} \mid \tilde{Y}_{\tilde{i},k} = 0) (1 - \tilde{p}_{ik}) + \eta \prod_{i=1}^{l} \sum_{\tilde{y}_{ik} = 0}^{1} P(Y_{\tilde{i},k} = y_{\tilde{i},k} \mid \tilde{Y}_{\tilde{i},k} = \tilde{y}_{\tilde{i},k}) P(\tilde{Y}_{\tilde{i},k} = \tilde{y}_{\tilde{i},k}) \right\}}.
\]

Derivation details are provided in Appendix A.
2.2.3 Halving

As its name suggests, halving works by first splitting a positive testing initial group into two equally (or as close to as possible) sized subgroups for retesting. Whenever a subgroup tests negative, all of its individuals are declared negative and no further splitting is performed. Whenever a subgroup tests positive, continued splitting occurs in the same manner until only individuals remain. Early origins of the halving protocol go back as far as Sobel and Groll (1959). More recently, halving and its close variants have been used in a number of infectious disease screening applications, including Litvak et al. (1994) and Priddy et al. (2007). Halving has even been described in the product literature for high throughput screening platforms (Tecan Group Ltd. 2007).

For a group of size $I_k = 2^s$, there are $s$ possible hierarchical splits that contain a particular individual specimen, where the last split results in individual testing. For practicality reasons, all possible hierarchical splits are rarely implemented. Instead, individual testing is performed on subgroups at a pre-determined $t^{th}$ split, where $t \leq s$. For this reason, we will only consider the $t = 2$ case, so that an individual can be tested at most three times.

To find $\omega_{ik}$ under halving, we continue to define $Z_k$ as the initial group response for group $k$, $k = 1, \ldots, K$. If the initial group tests positive ($Z_k = 1$), it is split into two subgroups that we denote by $k1$ and $k2$. The two subgroups are subsequently tested and provide the corresponding binary responses $Z_{k1}$ and $Z_{k2}$. If either subgroup tests positive, the third and final step is to individually test all members within a subgroup, where we continue to define $Y_{ik}$ as the individual retest outcome for individual $i$ from initial group $k$. To denote the true statuses
for the groups, subgroups, and individual tests under halving, we again use a
tilde over the respective letter symbol. We also continue to assume a constant
sensitivity and specificity for each test regardless of the group size.

For the halving protocol outlined above, there are five possible testing
scenarios involving the initial group and its two subgroups. These scenarios are:

1) \( Z_k = 0 \): Group \( k \) tests negative,
2) \( Z_k = 1, Z_{k1} = 0, Z_{k2} = 0 \): Group \( k \) tests positive, but both subgroups test
   negative,
3) \( Z_k = 1, Z_{k1} = 1, Z_{k2} = 0 \): Group \( k \) tests positive, subgroup \( k1 \) tests positive
   leading to individual testing for its members, and subgroup \( k2 \) tests
   negative,
4) \( Z_k = 1, Z_{k1} = 0, Z_{k2} = 1 \): Group \( k \) tests positive, subgroup \( k1 \) tests negative,
   and subgroup \( k2 \) tests positive leading to individual testing for its
   members,
5) \( Z_k = 1, Z_{k1} = 1, Z_{k2} = 1 \): Group \( k \) tests positive and both subgroups test
   positive leading to individual testing for members of both subgroups.

In Table 2.1, we provide expressions for \( \omega_{ik} \) in each of these scenarios. Derivations
are similar to those given in Section 2.2.2, but they are much more tedious due to
the additional split in the testing process. We present the derivations in
Appendix B.

### 2.2.4 Array Testing

Both Sections 2.2.2 and 2.2.3 describe protocols where individuals are initially
tested in non-overlapping groups. Phatarfod and Sudbury (1994) proposed a
fundamentally different protocol where specimens are arranged into a two-
dimensional array. Samples from specimens are combined within rows and within columns so that each individual is tested twice in overlapping groups. Specimens lying outside of any positive rows and columns are classified as negative. Specimens lying inside a positive row and/or column are potentially positive. This protocol is known as array (matrix) testing, and it is widely applied in high throughput screening applications, such as infectious disease testing (Tilghman et al. 2011), DNA screening (Berger et al. 2000), and systems biology (Thierry-Mieg 2006).

Because individuals are initially tested within one row and one column, we must modify our notation to reflect this. Define \( \tilde{Y}_{ij} \) as the true binary status (0 denotes negative, 1 denotes positive) for the individual whose specimen is located within row \( i \) and column \( j \), for \( i = 1, \ldots, I \) and \( j = 1, \ldots, J \). With this slight change in notation, our group testing regression model now can be rewritten as

\[
f(\tilde{\mu}_{ij}) = \beta_0 + \beta_1 x_{1,ij} + \ldots + \beta_{p-1} x_{p-1,ij},
\]

where the \( \tilde{Y}_{ij} \) are independent Bernoulli(\( \tilde{\mu}_{ij} \)) random variables, and the full-data log-likelihood function can be rewritten as

\[
\log[L(\beta)] = \sum_{i=1}^{I} \sum_{j=1}^{J} \tilde{y}_{ij} \log(\tilde{\mu}_{ij}) + (1 - \tilde{y}_{ij}) \log(1 - \tilde{\mu}_{ij}),
\]

if the true individual statuses were observed. In most screening applications, there will be more than \( IJ \) individuals, so more than one array will be needed. In those cases, we could add a third subscript to \( \tilde{Y}_{ij} \) to denote the array and include a third sum over the arrays in \( \log[L(\beta)] \). We avoid doing this for brevity.

As before, because the individual statuses are not observed directly, the EM algorithm is used to fit the regression model. Define \( R = (R_1, \ldots, R_I)' \) and \( C = (C_1, \ldots, C_J)' \) as vectors of row and column binary responses, respectively, for one
array. If identification of positive individuals is of interest, specimens lying at the intersections of positive rows and columns are retested individually. Additionally, specimens in positive testing rows without any positive testing columns in the array, which can occur when there is testing error, should be retested as well. The same is true when columns test positive without any rows testing positive.

Without loss of generality, we denote the collection of all potentially positive individual responses by \( Y_Q = (Y_{ij})_{(i,j)\in Q} \) where \( Q \) is the index set pertaining to the individual tests, that is

\[
Q = \{(s,t) \mid R_s = 1, C_t = 1, 1 \leq s \leq I, 1 \leq t \leq J \\
or R_s = 1, C_t = \cdots = C_J = 0, 1 \leq s \leq I \\
or R_s = \cdots = R_t = 0, C_t = 1, 1 \leq t \leq J \}.
\]

If there are no individual tests performed at all, we simply let \( Q = \emptyset \), the empty set.

Using all available test responses, we need to obtain the conditional expected value \( \omega_{ij} \equiv E(\tilde{Y}_{ij} \mid \mathcal{I}) \); however, when array testing is used as described above, there is no longer a closed form expression for it. Therefore, as suggested by Xie (2001), we implement a Gibbs sampling approach to estimate \( \omega_{ij} \). This involves successive sampling from the univariate conditional distribution of \( \tilde{Y}_{ij} \) given \( R = r, C = c, Y_Q = y_Q, \) and all of the other true individual binary statuses in the array, and this sampling is performed for each \( i \) and \( j \). After a large set of samples, all of the simulated \( \tilde{y}_{ij} \) values for each \( i \) and \( j \) can be averaged to find an estimate of \( \omega_{ij} \). Implementation details are described next.

For a given row and column combination \((i, j)\), define \( \tilde{Y}_{i,j} = \{\tilde{Y}_{i',j'} : i' = 1, \ldots, I, j' = 1, \ldots, J, (i', j') \neq (i, j)\} \); i.e., all possible true individual statuses excluding \( \tilde{Y}_{ij} \). The conditional distribution for \( \tilde{Y}_{ij} \mid \tilde{y}_{i,-j}, r, c, y_Q \) is
Bernoulli$(\gamma_{ij})$, where
\[
\gamma_{ij} \equiv P(\tilde{Y}_{ij} = 1 \mid \tilde{Y}_{-i,-j} = \tilde{y}_{-i,-j}, \mathbf{R} = \mathbf{r}, C = \mathbf{c}, Y_Q = y_Q)
\]
which we derive in Appendix C. With these conditional distributions, we generate samples \(\tilde{y}_{i1}^{(b)}, \ldots, \tilde{y}_{iJ}^{(b)}\) for \(b = 1, \ldots, B\), using the most updated \(\tilde{y}_{-i,-j}\). The estimate for \(\omega_{ij}\) is then taken to be \(\hat{\omega}_{ij} = (B - a)^{-1} \sum_{b=a+1}^{B} \tilde{y}_{ij}^{(b)}\), where \(a\) is a sufficiently large number of burn-in samples. The EM algorithm proceeds as usual where \(\hat{\omega}_{ij}\) replaces \(\omega_{ij}\) in each E-step. The negative information matrix can be estimated using these \(B\) Gibbs samples (e.g., see Xie (2001, p. 1961)).

2.3 Simulation Study

We use simulation to evaluate the regression estimates resulting from the group testing protocols described in Section 2.2. To begin, we consider the model
\[
\logit(p_{ik}) = \beta_0 + \beta_1 x_{ik},
\]
which is equivalently \(\logit(p_{ij}) = \beta_0 + \beta_1 x_{ij}\) for the array testing protocols. We let \(\beta_0 = -7\) and \(\beta_1 = 0.1\) and simulate covariates from a gamma(17, 1.4) distribution. The regression parameters and covariate distribution are chosen to produce a realistic group testing setting where most individuals have low risks of being positive and a few individuals having higher risks. Appendix D provides a histogram of the true individual probabilities for one simulated data set under these settings. Note that the overall mean prevalence is approximately 0.01.

Based on the logit model, we obtain the true probability of positivity \(p_{ik}\) (\(\tilde{p}_{ij}\) for array testing), which in turn is used to simulate a true individual status \(\tilde{Y}_{ik}\) (\(\tilde{Y}_{ij}\) for array testing). Individuals are then randomly assigned to groups of size \(I\) (\(I \times I\) arrays are used for array testing). Group, subgroup, and individual test responses for each protocol are simulated next by using \(\eta\) and \(\delta\) as Bernoulli
success probabilities. Group testing regression models are fit to these resulting responses. For comparison purposes, we also fit a model to individual testing data when testing error is present using the methodology outlined in Neuhaus (1999). We repeat the same simulation process for each simulated data set of size 5000 individuals. Large sample sizes such as this are common in group testing applications, including the example in Section 2.4.

2.3.1 Estimator Accuracy and Variance Estimation

Table 2.2 presents results on the accuracy of the parameter estimators and their standard errors for group sizes $I = 4, 12, 20$ and $\eta = \delta = 0.99$. The mean rows give each regression parameter estimate averaged over 1000 simulated data sets. The SE/SD rows examine the accuracy of the standard error estimates, where SD denotes the sample standard deviation of estimates across all simulated data sets, and SE denotes the corresponding averaged standard errors. Thus, a SE/SD ratio close to 1 suggests that the true standard errors are being estimated correctly. Note that because Gibbs sampling is used for array testing, the EM algorithm is much slower, so our array testing results are based on 300 simulated data sets.

We see from Table 2.2 that using the non-overlapping initial groups (IG; Section 2.2.1) results in comparatively poor estimates of the parameters and their standard errors. These estimates and standard errors become increasingly worse as the group size grows. In contrast, all of the other protocols perform similarly to individual testing, where averaged parameter estimates are close to corresponding true values and SE/SD ratios are close to 1. As these results show, there are important benefits from including retesting information from the Dorfman and halving protocols.
2.3.2 Improvements in Variance Estimation from Including Retests

As the results in Section 2.3.1 demonstrate, parameters and their corresponding standard errors can be estimated well when retests are included. In this subsection, we investigate directly the benefits of including retest information and how this extra information affects the slope estimator precision. Define the relative efficiency for $\hat{\beta}_1$ as

$$\text{RE}(\hat{\beta}_{1,\text{Retest}} \text{ to } \hat{\beta}_{1,\text{No retest}}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\text{Var}(\hat{\beta}_{1,b,\text{Retest}})}{\text{Var}(\hat{\beta}_{1,b,\text{No retest}})},$$

where $B$ denotes the number of simulations, $\hat{\beta}_{1,b,\text{Retest}}$ denotes the estimator for $\beta_1$ when retests are included in the $b^{th}$ simulated data set, and $\hat{\beta}_{1,b,\text{No retest}}$ is defined similarly when retests are not included. Note that we use the true variances in Equation (6), rather than estimated variances, due to the length of time it takes to fit a model for array testing. For Dorfman and halving, we compare their variances to IG. For array testing, we compare variances with and without retests.

Figure 2.1 displays the relative efficiencies from $B = 500$ new simulated data sets for group sizes $I = 4, 6, \ldots, 20$ when $\eta = \delta = 0.99$ and $\eta = \delta = 0.95$. Overall, we see very large efficiency gains from including retesting information. Using retests with array testing provides the smallest gain (but still noteworthy), which is likely due to each individual already being part of two groups even if no retests are performed. Halving results in larger gains than Dorfman, where the differences between them are more pronounced for smaller $\eta$ and $\delta$. This occurs because halving generally will always result in a lower classification error rate than Dorfman (e.g., see Black et al. 2012), which then leads to less uncertainty in the parameter estimates under halving. Overall, the efficiencies for all protocols
grow as the group size does. This is explained by the fact that protocols without retests observe less information as the group size increases. In contrast, retesting will moderate the amount of information lost for larger group sizes.

Figure 2.2 provides plots of the averaged $\text{Var}(\hat{\beta}_1)$ for all simulated data sets. One will note that the averaged $\text{Var}(\hat{\beta}_1)$ for the two testing protocols without retests increases as the group size increases. This is similar to Figure 2.2 where it was shown that $\text{RE}(\hat{\beta}_{1,\text{Retest}} \text{ to } \hat{\beta}_{1,\text{No retest}})$ increases as a function of the group size. Conversely, when retests are included in a protocol, the averaged $\text{Var}(\hat{\beta}_1)$ changes very little across the group sizes because positive individuals are still identified (subject to testing error).

Ordered by their averaged $\text{Var}(\hat{\beta}_1)$, we can informally write Dorfman $> \text{halving} >$ array testing with retests. Interestingly, each of these protocols (and also array testing without retests for smaller sensitivity, specificity, and group size levels) has a smaller variance than that found through individual testing, while also resulting in a smaller number of tests (see Appendix E). In other words, not only do these protocols have the potential to drastically reduce the costs needed for classification, but these protocols provide better regression estimates! Note that Liu et al. (2012, Theorem 2) has recently observed this same phenomenon in the absence of covariates. Through additional simulations (not shown), we have seen that the gains from group testing in estimation efficiency (over individual testing) do diminish as the assay sensitivity and specificity both approach 1. This is an expected result because both individual and group testing are likely to find all positive and negative individuals when assays are perfect or nearly perfect.
2.3.3 Average Number of Tests per Unit of Information

Each protocol uses a different number of tests to estimate the regression parameters. To take this aspect into account, we define the average number of tests per unit of information for $\beta_1$ to be

$$\psi = \frac{1}{B} \sum_{b=1}^{B} \frac{n_b}{1 / \text{Var}(\hat{\beta}_{1,b})} = \frac{1}{B} \sum_{b=1}^{B} n_b \text{Var}(\hat{\beta}_{1,b}),$$

where $n_b$ is the total number of tests performed for a protocol and $\hat{\beta}_{1,b}$ is the estimated $\beta_1$ for the $b^{th}$ simulated data set. The smaller that $\psi$ is, the fewer the number of tests are needed comparatively to obtain the same amount of information about $\beta_1$. A similar measure was used by Chen and Swallow (1990, p. 1037) when evaluating the benefits of retesting for overall prevalence estimation.

Figure 2.3 plots values of $\psi$ for all group testing protocols for the same simulations as in Section 2.3.2. Individual testing results in $\psi = 3.53$ for $\eta = \delta = 0.99$ and $\psi = 8.80$ for $\eta = \delta = 0.95$; these values were excluded from the figure to avoid distorting the plots. Comparing between the plots, we see that $\psi$ is larger for $\eta = \delta = 0.95$ than for $\eta = \delta = 0.99$, which is a byproduct of increased uncertainty when $\eta$ and $\delta$ are smaller. Within each plot, we again see the benefits of including retests in the estimation process. Dorfman, halving, and array testing with retests have smaller $\psi$ values than their corresponding protocols that do not include retests. Among those that include retests, halving always provides a smaller $\psi$ than Dorfman’s protocol. Also, array testing with retests provides values of $\psi$ close to that of halving for larger group sizes.
2.3.4 Additional Simulations

To determine if our findings in this section remain in other situations, we have performed a number of additional simulations. These simulations include using a different covariate distribution and different regression parameter values, which also allows us to examine different overall prevalence levels. In summary, we have found that the same conclusions hold in these other situations, and some of these results are included in Appendix F.

2.4 Infertility Prevention Project

The purpose of the Infertility Prevention Project (IPP) in the United States is to prevent complications from chlamydia and gonorrhea infections that lead to infertility. Annually, over 3 million screenings for these infections are reported to the IPP program. Due to the large number of tests, some states, including Idaho and Iowa, already use group testing to reduce costs. For this dissertation, we will examine data from Nebraska, where individual testing is performed at the Nebraska Public Health Laboratory (NPHL) for the entire state. In order to reduce costs, the laboratory has an interest in adopting group testing – not only to reduce the number of tests, but also to estimate risk factor specific probabilities of infection. Thus, both case identification and estimation are important goals for the NPHL.

We focus on the 6,139 test results from males who had their urine tested for chlamydia in 2009. To examine how group testing would have worked with these individuals, we artificially construct group, subgroup, and individual retest responses for each group testing protocol by treating the known individual test
results as the true statuses. Each test response is simulated by taking into account assay sensitivity and specificity at the NPHL ($\eta = 0.93$, $\delta = 0.95$). Initial groups are formed chronologically based on when specimens arrived at the NPHL. The “optimal” sizes for these initial groups are found by minimizing the expected number of tests (e.g., see Kim et al. (2007) for expected value formulas) as a function of the 2008 overall prevalence of 0.077. These optimal group sizes are 5 for IG and Dorfman and 8 for halving and array testing.

A first-order logit regression model is fit to the responses from each protocol with the following covariates: age, race (represented by three indicator variables), symptoms, urethritis, and risk history variables (multiple partners, new partner in the last 90 days, contact with someone who has a sexually transmitted disease). All covariates are dichotomous (0 and 1) except for age. For comparison purposes, we again fit the same model to the original individual responses while incorporating testing errors using the methodology of Neuhaus (1999).

Table 2.3 gives the parameter estimates from all fitted models and the number of tests required for each protocol. Overall, all estimates are close to each other for the same corresponding covariates. Each group testing protocol that includes retests has smaller standard errors than those for the individual testing model, consistent with our findings in Section 2.3. Using a level of significance of 0.05 with Wald tests, individual testing and group testing protocols with retests agree on the same set of important covariates. These results illustrate the potential advantages of using group testing at the NPHL—both in terms of estimation and the resulting large-sample inference, but also because of the opportunity to drastically reduce the number of tests needed. For example, halving requires 2898 tests overall, which is a 52.7% reduction from individual testing. Even the
simpler Dorfman protocol requires only 3458 tests overall, a 43.7% reduction in tests when compared to individual testing.

2.5 Discussion

In this chapter, we have outlined how to estimate a group testing regression model when retesting information is available from three commonly used protocols. Functions to fit the models are available in R’s binGroup package (Bilder et al. 2010). Including retests leads to large reductions in estimator variability while also improving estimator accuracy. Overall, halving and array testing with retests are the best protocols when taking into account the number of tests as well as the estimator variability. We also showed that group testing can lead to more efficient estimates of regression parameters than individual testing. This is an extremely important finding, because it shows that more information can be gained from a statistical analysis by actually doing less in terms of testing.

Group size selection is an important consideration in most applications where group testing is used (e.g., see Swallow (1985)). Aside from assay considerations, the optimal group size is the one that leads to the smallest number of tests while still providing as much information as possible. Our research shows the average number of tests per unit of information stays relatively stable over a large range of group sizes when retests are included. Thus, protocols with retests are somewhat robust to the group size used, which makes its choice not as critical as when retesting is omitted.

The EM algorithm proposed by Xie (2001) can be used to fit models for data arising from any group testing protocol. While our paper focused on three
commonly used protocols for case identification, other protocols exist. In particular, array testing can be implemented with a master group for each array and/or in more than two dimensions (Kim et al. 2007; Kim and Hudgens 2009). Future research could examine these other protocols to determine if more estimation benefits result from their implementation. In the case of array testing, all protocols will likely need to use the Gibbs sampling approach outlined in Section 2.2.4 to estimate a conditional expectation for every cell within an array. This can be time consuming depending on the size of the arrays and how many arrays there are. Potentially, parallel processing could be used with one core processor per array to reduce the model fitting time.

2.6 References


Introduction of *Chlamydia Trachomatis* screening for young women in Germany. *Journal der Deutschen Dermatologischen Gesellschaft* 6, 1032-1037.


Table 2.1. The numerator and denominator for $\omega_{ik}$ for the halving protocol in Section 2.2.3. To simplify the expressions, $	ilde{q}_{ik} = 1 - \hat{p}_{ik}$ and $\lambda_{ik}^{(y)} = P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik})$ are used, and we assume individual $i$ is within subgroup $k1$.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Numerator</th>
<th>Denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(1 - \eta)\hat{p}_{ik}$</td>
<td>$1 - \theta_{ik}$</td>
</tr>
<tr>
<td>2</td>
<td>$\eta(1 - \eta)\hat{p}<em>{ik} \left[ \delta \prod</em>{i \in k1} \tilde{q}<em>{ik} + (1 - \eta) \left( 1 - \prod</em>{i \in k1} \tilde{q}_{ik} \right) \right]$</td>
<td>$\delta(1 - \delta) \prod_{i \in k1} \tilde{q}<em>{ik} + \delta \eta(1 - \eta) \left( \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) \left( 1 - \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) + \delta \eta(1 - \eta) \left( 1 - \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) \left( \prod</em>{i \in k2} \tilde{q}<em>{ik} \right) + \eta(1 - \eta)^2 \left( 1 - \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) \left( 1 - \prod</em>{i \in k2} \tilde{q}_{ik} \right)$</td>
</tr>
<tr>
<td>3</td>
<td>$\eta^2 \hat{p}<em>{ik} \lambda</em>{ik}^{(1)} \sum_{y_{ik} \neq 0} \left( \prod_{i \in k1, y_{ik} \neq 0} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right)$</td>
<td>$\delta(1 - \delta)^2 \left( \prod_{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) \left( \prod</em>{i \in k2} \tilde{q}<em>{ik} \right) + (1 - \delta) \eta(1 - \eta) \left( \prod</em>{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) \left( 1 - \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) + \delta \eta^2 \sum</em>{y_{ik} = 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( \prod_{i \in k2} \tilde{q}<em>{ik} \right) + \eta^2 (1 - \eta) \sum</em>{y_{ik} \neq 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( 1 - \prod_{i \in k2} \tilde{q}_{ik} \right)$</td>
</tr>
<tr>
<td>4</td>
<td>$\eta(1 - \eta)\hat{p}<em>{ik} \left[ \sum</em>{i \in k1, y_{ik} = 0} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right]$</td>
<td>$\delta(1 - \delta)^2 \left( \prod_{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) \left( \prod</em>{i \in k2} \tilde{q}<em>{ik} \right) + (1 - \delta) \eta(1 - \eta) \left( \prod</em>{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) \left( 1 - \prod</em>{i \in k1} \tilde{q}<em>{ik} \right) + \delta \eta^2 \sum</em>{y_{ik} = 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( \prod_{i \in k2} \tilde{q}<em>{ik} \right) + \eta^2 (1 - \eta) \sum</em>{y_{ik} \neq 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( 1 - \prod_{i \in k2} \tilde{q}_{ik} \right)$</td>
</tr>
<tr>
<td>5</td>
<td>$\eta^2 \hat{p}<em>{ik} \lambda</em>{ik}^{(1)} \sum_{y_{ik} \neq 0} \left( \prod_{i \in k1, y_{ik} \neq 0} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right)$</td>
<td>$\delta(1 - \delta)^3 \left( \prod_{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) + (1 - \delta) \eta^2 \left( \prod</em>{i \in k1} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) \sum</em>{y_{ik} = 0} \left( \prod_{i \in k2} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) + (1 - \delta) \eta^2 \sum_{y_{ik} = 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( \prod_{i \in k2} \lambda_{ik}^{(0)} \tilde{q}<em>{ik} \right) + \eta^3 \sum</em>{y_{ik} \neq 0} \left( \prod_{i \in k1} \lambda_{ik}^{(y_{ik})} P(\tilde{Y}<em>{ik} = \tilde{y}</em>{ik}) \right) \left( \prod_{i \in k2} \lambda_{ik}^{(0)} \tilde{q}_{ik} \right)$</td>
</tr>
</tbody>
</table>
Table 2.2. Parameter estimates and their standard errors based on 1000 (300 for array testing) simulated data sets with $\beta_0 = -7$, $\beta_1 = 0.1$, and $\eta = \delta = 0.99$. The mean row includes the averaged estimate across all simulated data sets. The SE/SD row gives the averaged standard error over all simulated data sets (SE) divided by the sample standard deviation of the estimates across all data sets (SD).

<table>
<thead>
<tr>
<th>Protocol</th>
<th>$I = 4$</th>
<th>$I = 12$</th>
<th>$I = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_0$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_0$</td>
</tr>
<tr>
<td>Individual Mean</td>
<td>-7.003 0.099</td>
<td>-7.013 0.099</td>
<td>-7.016 0.099</td>
</tr>
<tr>
<td>SE/SD</td>
<td>0.983 0.977</td>
<td>0.970 0.966</td>
<td>1.002 0.987</td>
</tr>
<tr>
<td>IG Mean</td>
<td>-6.918 0.096</td>
<td>-6.840 0.091</td>
<td>-6.628 0.081</td>
</tr>
<tr>
<td>SE/SD</td>
<td>0.961 0.948</td>
<td>0.886 0.854</td>
<td>0.861 0.840</td>
</tr>
<tr>
<td>Dorfman Mean</td>
<td>-6.995 0.099</td>
<td>-7.013 0.100</td>
<td>-6.983 0.099</td>
</tr>
<tr>
<td>SE/SD</td>
<td>1.002 1.008</td>
<td>0.982 0.982</td>
<td>0.978 0.980</td>
</tr>
<tr>
<td>Halving Mean</td>
<td>-7.000 0.099</td>
<td>-7.015 0.099</td>
<td>-7.021 0.098</td>
</tr>
<tr>
<td>SE/SD</td>
<td>1.016 1.020</td>
<td>0.982 0.982</td>
<td>0.978 0.973</td>
</tr>
<tr>
<td>Array w/o retesting Mean</td>
<td>-7.024 0.099</td>
<td>-6.984 0.099</td>
<td>-7.023 0.099</td>
</tr>
<tr>
<td>SE/SD</td>
<td>1.007 1.044</td>
<td>0.981 0.997</td>
<td>0.989 0.991</td>
</tr>
<tr>
<td>Array w/ retesting Mean</td>
<td>-7.022 0.100</td>
<td>-7.010 0.100</td>
<td>-7.018 0.099</td>
</tr>
<tr>
<td>SE/SD</td>
<td>0.982 1.017</td>
<td>1.001 1.011</td>
<td>0.979 0.979</td>
</tr>
</tbody>
</table>
Table 2.3. Parameter estimates and estimated standard errors for the chlamydia screening data. The “p-value” column gives Wald test p-values for testing whether or not a regression parameter is equal to 0. Note that an overall test is performed for all levels of the variable Race. The number of tests performed by each protocol is in parenthesis after the protocol name.

<table>
<thead>
<tr>
<th>Term</th>
<th>Individual (6139)</th>
<th>IG (1228)</th>
<th>Dorfman (3458)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>p-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.46</td>
<td>0.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.03</td>
<td>0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race level #1</td>
<td>0.79</td>
<td>0.15</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race level #2</td>
<td>0.80</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>Race level #3</td>
<td>0.44</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>0.45</td>
<td>0.16</td>
<td>0.004</td>
</tr>
<tr>
<td>Urethritis</td>
<td>1.29</td>
<td>0.33</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Multiple partners</td>
<td>0.44</td>
<td>0.19</td>
<td>0.019</td>
</tr>
<tr>
<td>New partner</td>
<td>0.17</td>
<td>0.20</td>
<td>0.407</td>
</tr>
<tr>
<td>Contact to a STD</td>
<td>1.04</td>
<td>0.15</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Halving (2898)</th>
<th>Array w/o retesting (1541)</th>
<th>Array w/ retesting (3097)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>p-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.39</td>
<td>0.22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.04</td>
<td>0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race level #1</td>
<td>0.64</td>
<td>0.14</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race level #2</td>
<td>0.47</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Race level #3</td>
<td>0.68</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>0.63</td>
<td>0.15</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Urethritis</td>
<td>1.07</td>
<td>0.34</td>
<td>0.002</td>
</tr>
<tr>
<td>Multiple partners</td>
<td>0.35</td>
<td>0.16</td>
<td>0.029</td>
</tr>
<tr>
<td>New partner</td>
<td>0.11</td>
<td>0.20</td>
<td>0.600</td>
</tr>
<tr>
<td>Contact to a STD</td>
<td>1.16</td>
<td>0.15</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Figure 2.1. Relative efficiencies calculated by Equation (6) based on 500 simulated data sets. Dorfman and halving protocols are compared to IG. Array testing is compared with and without retests.
Figure 2.2. Averaged $\text{Var}(\hat{\beta})$ for 500 simulated data sets. The dashed horizontal line corresponds to $\text{Var}(\hat{\beta})$ from individual testing. The right-side plots are the same as those on the left-side except we omit IG in order to reduce the y-axis scale.
Figure 2.3. Average number of tests per unit of information calculated by Equation (7) based on 500 simulated data sets.
Chapter 3: Paper #2 - Regression Analysis for Multiple-Disease Group Testing Data

Abstract

Group testing, where groups of individual specimens are composited to test for the presence of a disease (or other binary trait), is a procedure commonly used to reduce the costs of screening a large number of individuals. Group testing data are unique in that only group responses may be observed, but inferences are needed at the individual level. A further methodological challenge arises when individuals are tested in groups for multiple diseases simultaneously, because the unobserved individual disease statuses are likely to be correlated. In this paper, we propose the first regression techniques for multiple-disease group testing data. We develop an expectation-solution based algorithm that provides consistent parameter estimates and natural large-sample inference procedures. Our proposed methodology is applied to chlamydia and gonorrhea screening data collected in Nebraska as part of the Infertility Prevention Project.

KEY WORDS: Correlated binary data; Expectation-solution algorithm; Generalized estimating equations; Latent response; Pooled testing; Unobserved response.
3.1 Introduction

Researchers are often interested in modeling the disease infection status of individuals to identify important risk factors and to estimate subject-specific risk probabilities. In many cases, pooling specimens (e.g., blood, urine, swabs, etc.) through group testing offers a novel approach to significantly reduce the number of tests, the time expended, and the overall costs. These practical benefits have led to the adoption of group testing in a number of infectious disease applications, including blood donation screening by the American Red Cross (American Red Cross, 2012), opportunistic chlamydia and gonorrhea testing in medical clinics (Gaydos, 2005), and Bovine Viral Diarrhea virus detection for the cattle industry (Munoz-Zanzi et al., 2006). Group testing has also proven to be beneficial in other areas including pharmaceutical drug discovery (Remlinger et al., 2006), plant pathology (Tebbs and Bilder, 2004), genotyping (Chi et al., 2009), and food contamination testing (Fahey, Ourisson, and Degnan, 2006).

Statistical research in group testing has traditionally focused on estimating the prevalence of disease in a homogeneous population. More recently, research has shifted towards incorporating individual covariate information to produce individual-specific estimates in a regression context. Vansteelandt, Goetghebeur, and Verstraeten (2000) and Xie (2001) are commonly regarded as the seminal papers in this area. Vansteelandt et al. (2000) provides a generalized linear model regression approach that uses only the initial group responses for estimation. Xie’s (2001) approach is more flexible by allowing for different classes of regression models and the inclusion of additional information from retesting subsets of positive groups. Several very recent papers have expanded on the work
of Vansteelandt et al. (2000) and Xie (2001). Specifically, Bilder and Tebbs (2009) provide a thorough comparison of individual and group testing regression model estimates, Chen, Tebbs, and Bilder (2009) examine mixed-effects models, Delaigle and Meister (2011) and Delaigle and Hall (2012) formulate a nonparametric modeling approach. Group testing regression models even have been used to diagnose model misspecification with individual response data, as illustrated by Huang (2009).

When viewed collectively, research in group testing regression modeling has one notable shortcoming; namely, the available methodology involves only single-disease models. However, in many screening applications, testing is performed not for one disease but for multiple diseases at the same time – often using the same assay. For example, the American Red Cross uses group testing to screen millions of blood donations per year for HIV, hepatitis B, and hepatitis C with a single assay (Stramer et al., 2004; American Red Cross, 2012). Also, as part of the nationally implemented Infertility Prevention Project, the Nebraska Public Health Laboratory (NPHL) uses the GenProbe Aptima Combo 2 assay to test thousands of individual specimens each year for chlamydia and gonorrhea simultaneously. Despite the ubiquity of multiple-disease screening in practice, Hughes-Oliver and Rosenberger (2000) is the only paper that has addressed the multiple-disease problem in the group testing literature, and they do so assuming that the population is homogeneous and that diagnostic tests are perfect.

The purpose of our paper is to develop new group testing regression models for multiple-disease screening data in heterogeneous populations with imperfect diagnostic tests. In essence, our research deals with modeling correlated binary
data, but with the unique aspect that the underlying disease responses are unobserved for each individual. Broadly speaking, our paper can be viewed as a generalization of Vansteelandt et al. (2000) and Xie (2001) to model multiple-disease statuses and, at the same time, a generalization of Hughes-Oliver and Rosenberger (2000) to incorporate covariate information and imperfect diagnostic tests.

The remainder of this paper is organized as follows. Section 3.2 describes the notation and states the model of interest. Section 3.3 shows how the expectation-solution (ES) algorithm of Elashoff and Ryan (2004) can be used to model multiple-disease statuses with group testing responses. Due to the complicated relationship between the unobserved individual and observed group responses when diagnostic testing error is present, we develop new ways to approximate the true correlation structure among the unobserved individuals. Section 3.4 presents simulation evidence demonstrating that parameter estimates are consistent and that Wald confidence intervals achieve their stated confidence levels in realistic settings. Section 3.5 applies this work to the chlamydia and gonorrhea screening data from the NPHL. Finally, Section 3.6 summarizes this work and suggests future areas of research.

3.2 Notation and Model

Define $\tilde{Y}_{ijk}$ as the true unknown individual status of disease $j$ for the $i^{th}$ individual in group $k$, where $i = 1, ..., I$, $j = 1, ..., J$, $k = 1, ..., K$, and suppose that these random variables are independent across $i$ and $k$. The value for $\tilde{Y}_{ijk}$ is 0 for a negative response and 1 for a positive response; we use this standard convention
for all subsequently defined binary random variables. For each individual $i$, \( \tilde{Y}_{ik} = (\tilde{Y}_{i1k}, ..., \tilde{Y}_{iJk})' \) contains $J$ unobserved disease statuses that are likely correlated.

Define $Z_{jk}$ as the observed group binary response for the $j^{th}$ disease and the $k^{th}$ group. We assume that all groups are non-overlapping and that each individual is within one group. If group tests are perfectly accurate, as assumed in Hughes-Oliver and Rosenberger (2000), $Z_{jk} = 1$ if and only if $\sum_{i=1}^{I_k} \tilde{Y}_{ijk} > 0$ and $Z_{jk} = 0$ if and only if $\sum_{i=1}^{I_k} \tilde{Y}_{ijk} = 0$. Of course, assays are unlikely to be perfect in practice, so one must account for this uncertainty. For disease $j$, define the group test sensitivity and specificity as $\eta_j = P(Z_{jk} = 1 \mid \tilde{Z}_{jk} = 1)$ and $\delta_j = P(Z_{jk} = 0 \mid \tilde{Z}_{jk} = 0)$, respectively, where $\tilde{Z}_{jk}$ denotes the true group binary response for disease $j$ and group $k$. We assume the sensitivity and specificity are known for each disease and are not dependent on pool size; these assumptions are analogous to those made by Vansteelandt et al. (2000) and Xie (2001) for single-disease group testing regression models and by Neuhaus (1999, 2002) for individual testing regression models. We can then express $\theta_{jk} = P(Z_{jk} = 1)$ in terms of the true individual probabilities $P(\tilde{Y}_{ijk} = 1) \equiv \tilde{p}_{ijk}$ as

\[
\theta_{jk} = P(Z_{jk} = 1 \mid \tilde{Z}_{jk} = 1)P(\tilde{Z}_{jk} = 1) + P(Z_{jk} = 1 \mid \tilde{Z}_{jk} = 0)P(\tilde{Z}_{jk} = 0) \\
= \eta_j + (1 - \delta_j - \eta_j)\prod_{i=1}^{I_k} (1 - \tilde{p}_{ijk}).
\]

(8)

With covariates $x_{ik} = (x_{i1k}, ..., x_{I-1,ik})'$ collected on each individual, our goal is to estimate $\tilde{p}_{ijk}$ when only the observed group responses $Z_{jk}$ are available, similar to Vansteelandt et al. (2000) with single-disease models. In all subsequent expectations written in this chapter, we condition on the full set of covariates $x_{ik}$.
as we did for $\tilde{p}_{ijk}$, but we suppress this specification for notational simplicity. We consider models of the form

$$f(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{ij}x_{1ik} + \cdots + \beta_{p-1,j}x_{p-1,ik},$$

(9)

where $f(\cdot)$ is a known monotonic, differentiable function and $\beta_{rj}$ ($r = 0, \ldots, p - 1, j = 1, \ldots, J$) is a regression parameter. Using a joint model, as in Equation (9), not only enables one to analyze group testing data as they naturally arise from multiple-disease screening assays, but it also allows one to incorporate the within-individual correlation across the $J$ diseases. We demonstrate in Section 3.4 that our joint modeling approach in realistic settings provides more efficient regression estimators than using $J$ separate single-disease group testing models. This is because separate modeling discards important information about how the $J$ disease statuses are related.

### 3.3 Expectation-Solution Algorithm

The ES algorithm, introduced by Elashoff and Ryan (2004), is a generalization of the expectation-maximization (EM) algorithm given by Dempster et al. (1977). The algorithm iterates between two steps: the E-step, which computes the expectation of the complete data given the observed data, and the S-step, which substitutes the expected values into the complete-data estimating equations and solves the equations for the model parameters. The generalization given in Elashoff and Ryan (2004) allows these estimating equations to take on a variety of forms, including generalized estimating equations. We utilize the ES algorithm by treating the unobserved individual responses in group testing as “missing” and modify the algorithm to estimate Equation (9) using the observed group
responses. Our application of the ES algorithm requires additional work to estimate the correlation among the unobserved individual responses, as shown in Section 3.3.2.

### 3.3.1 Estimating Equations

To explain our model fitting approach, consider the hypothetical situation where the true individual responses $\tilde{Y}_{ijk}$ are observed and standard generalized estimating equation (GEE) methodology is used to estimate the model in Equation (9). Let $R(\alpha)$, where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_J)'$, denote the $J \times J$ working correlation matrix for the true individual responses. Define $\text{Cov}(\tilde{Y}_{ik}) = V_{ik} = B_{ik}^{1/2} R(\alpha) B_{ik}^{1/2}$ where $B_{ik} = \text{Diag}(\tilde{p}_{ijk}(1 - \tilde{p}_{ijk}))$. The estimating equations are

$$
\Psi(\beta, \alpha) = \sum_k \sum_i \Psi_{ik}(\beta, \alpha) = \sum_k \sum_i D_{ik}' V_{ik}^{-1} (\tilde{y}_{ik} - \tilde{p}_{ik}) = 0, \quad (10)
$$

where $\beta = (\beta_{01}, ..., \beta_{p-1,1}, \beta_{02}, ..., \beta_{p-1,2})'$, $D_{ik} = (\partial / \partial \beta) \tilde{p}_{ik}$, $\tilde{p}_{ik} = (\tilde{p}_{1ik}, ..., \tilde{p}_{pik})'$, $\tilde{y}_{ik}$ is a realization of $\tilde{Y}_{ik}$, $0$ is a $pJ \times 1$ vector of 0’s, and $\Psi_{ik}(\beta, \alpha) = D_{ik}' V_{ik}^{-1} (\tilde{y}_{ik} - \tilde{p}_{ik})$ is the contribution of the $i^{th}$ subject in the $k^{th}$ group to the estimating equations. If realizations of the individual responses $\tilde{Y}_{ijk}$ were available, parameter estimates would be found by successively estimating $\alpha$ and solving Equation (10) for $\beta$ in an iterative manner until convergence.

Because the individual responses $\tilde{Y}_{ijk}$ are not observed in group testing, we cannot use standard GEE methodology as stated above. However, analogous to the use of the expectation-maximization (EM) algorithm approach described in Xie (2001) for a single disease, we can replace the individual responses in Equation (10) by their expected values, conditional on the group responses $Z = (Z_{1k}, ..., \tilde{p}_{1ik}, ..., \tilde{p}_{pik})'$. 

...
Because a conditional expectation involving $\tilde{Y}_{ijk}$ depends only on its corresponding group response, it suffices to calculate $E(\tilde{Y}_{ijk} | Z_{jk} = 1) = \eta_i \tilde{p}_{ijk} / \theta_{jk}$ and $E(\tilde{Y}_{ijk} | Z_{jk} = 0) = (1 - \eta_i) \tilde{p}_{ijk} / (1 - \theta_{jk})$. Replacing $\tilde{y}_{ijk}$ with $E(\tilde{Y}_{ijk} | z_{jk})$, Equation (10) becomes

$$
\Psi^{\text{obs}}(\beta, \alpha) = \sum_k \sum_i \Psi^{\text{obs}}_{ik}(\beta, \alpha) = \sum_k \sum_i D_i \tilde{V}_{ik}^{-1}(\omega_{ik} - \tilde{p}_{ik}) = 0, \tag{11}
$$

where $\omega_{ik} = (E(\tilde{Y}_{i1k} | z_{1k}), ..., E(\tilde{Y}_{i|k} | z_{|k}))'$ and $\Psi^{\text{obs}}_{ik}(\beta, \alpha) = D_i \tilde{V}_{ik}^{-1}(\omega_{ik} - \tilde{p}_{ik})$.

The symbol $\Psi^{\text{obs}}$ indicates that Equation (11) no longer involves any unknown individual responses. The ES algorithm successively estimates $\alpha$ and solves Equation (11) for $\beta$ in an iterative manner to obtain parameter estimates. The initial estimate of $\beta$ can be found by estimating separate models for each disease with the methodology in Xie (2001). Note that the expectations $E(\tilde{Y}_{ijk} | z_{jk})$ are updated at each iteration to correspond to the current estimate of $\beta$. Estimating $\alpha$ at each iteration is not straightforward, so we discuss it thoroughly in the next subsection. The final iterative solution to Equation (11) at convergence is the estimate of $\beta$, which we denote by $\hat{\beta}$.

### 3.3.2 Correlation Estimation

To estimate $\alpha$, we need to first identify the relationship between $\text{Cov}(Z_{jk}, Z_{jk}')$, which we can estimate from the observed group responses, and $\text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{i'jk})$, which involves the unobserved individual responses. This relationship is given in the following theorem.

**THEOREM 1:** Assuming that the observed group responses are independent given the true group statuses, the covariance between $Z_{jk}$ and $Z_{jk}'$, when written as a function of the correlation of the unknown individual responses, is
\[ \text{Cov}(Z_{jk}, Z_{j'k}) = \Delta_{j'j} \prod_{i=1}^{l} \left\{ \text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{i'kj}) \sqrt{\text{Var}(\tilde{Y}_{ijk})} \text{Var}(\tilde{Y}_{i'kj}) + (1 - \hat{p}_{ijk})(1 - \hat{p}_{i'kj}) \right\} - \prod_{i=1}^{l} (1 - \hat{p}_{ijk})(1 - \hat{p}_{i'kj}) \]  
(12)

for \( 1 \leq j, j' \leq J \) and \( k = 1, \ldots, K \), where \( \Delta_{j'j} = (\delta_j + \eta_{jj} - 1)(\delta_{j'} + \eta_{jj'} - 1) \).

The proof of Theorem 1 is given in Appendix G. The importance of Theorem 1 is that it provides a convenient way to obtain method of moments estimates for \( \text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{i'kj}) \). Suppose an estimate of the model given in Equation (9) is available so that we can then estimate \( \theta_{jk} \), denoted by \( \hat{\theta}_{jk} \), through Equation (8). Define \( \hat{r}_{jk} = z_{jk} - \hat{\theta}_{jk} \) as residuals from the model’s fit, where \( z_{jk} \) is the realization of \( Z_{jk} \). After replacing \( \text{Cov}(Z_{jk}, Z_{j'k}) \) with \( \hat{r}_{jk} \hat{r}_{j'k} \) in the left-hand side of (12), we create one equation for each \( \alpha_s \) \((s = 1, \ldots, S)\). We argue in Appendix H that one unique solution \( \hat{\alpha}_s \) can be found in each equation and that \( \hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_s) \) is a consistent estimator of \( \alpha \) when \( \beta \) is known.

To illustrate this approach, suppose that there is an exchangeable correlation structure between \( \tilde{Y}_{ijk} \) and \( \tilde{Y}_{i'kj} \), say, \( \text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{i'kj}) = \alpha \), so that \( S = 1 \). The estimate of \( \alpha \) is obtained by solving

\[ \sum_{k=1}^{K} \sum_{j > j'} \hat{r}_{jk} \hat{r}_{j'k} = \sum_{k=1}^{K} \sum_{j > j'} \Delta_{j'j} \prod_{i=1}^{l} \left\{ \alpha \sqrt{\hat{p}_{ijk}(1 - \hat{p}_{ijk})} \hat{p}_{ij'k}(1 - \hat{p}_{ij'k}) + (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \right\} - \prod_{i=1}^{l} (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \]  
(13)

for \( \alpha \), where \( \hat{p}_{ijk} \) is the model’s estimate of \( \hat{p}_{ijk} \). Alternatively, if one specifies an unstructured correlation matrix so that \( \text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{i'kj}) = \alpha_{ij'} \), for \( j \neq j' \), we obtain \( J(J - 1)/2 \) equations of the form

\[ \sum_{k=1}^{K} \hat{r}_{jk} \hat{r}_{j'k} = \Delta_{j'j} \sum_{k=1}^{K} \prod_{i=1}^{l} \left\{ \alpha_{ij'} \sqrt{\hat{p}_{ijk}(1 - \hat{p}_{ijk})} \hat{p}_{ij'k}(1 - \hat{p}_{ij'k}) + (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \right\} - \prod_{i=1}^{l} (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \]  
(14)
The \((j, j')\)th element of \(\mathbf{R}(\alpha)\) can be estimated by solving Equation (14) for \(\alpha_{j,j'}\). Estimation for other correlation structures is performed in a similar manner.

Because \(\text{Cov}(Z_{jk}, Z_{j'k})\) is a polynomial function of \(\text{Corr}(Y_{ijk}, Y_{ij'k})\) of degree \(I_k\), obtaining the coefficients for this function can be computationally expensive when the group size \(I_k\) is large. Fortunately, we have found that the higher order (\(\geq 3\)) coefficients of \(\text{Corr}(Y_{ijk}, Y_{ij'k})\) are almost always negligible. As a result, it usually suffices to use the linear and quadratic terms to estimate \(\alpha\). For example, for an unstructured working correlation matrix, the linear and quadratic coefficients of \(\alpha_{j,j'}\) in (13) are computed as

\[
\hat{c}_{j,j',k} = \Delta_{j,j'} \left\{ \prod_{i=1}^{I_k} (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \right\} \sum_{i=1}^{I_k} \sqrt{\frac{\hat{p}_{ijk} \hat{p}_{ij'k}}{(1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k})}}
\]

and

\[
\hat{d}_{j,j',k} = \Delta_{j,j'} \left\{ \prod_{i=1}^{I_k} (1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k}) \right\} \times \sum_{1 \leq i < k \leq I_k} \sqrt{\frac{\hat{p}_{ijk} \hat{p}_{ij'k}}{(1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k})}} \sqrt{\frac{\hat{p}_{ijk} \hat{p}_{ij'k}}{(1 - \hat{p}_{ijk})(1 - \hat{p}_{ij'k})}}
\]

respectively. The estimate \(\hat{\alpha}_{j,j'}\) solves

\[
\sum_{k=1}^{K} \hat{r}_{jk} \hat{r}_{j'k} = \sum_{k=1}^{K} \hat{c}_{j,j',k} \hat{\alpha}_{j,j'}
\]

using a first-order approximation or

\[
\sum_{k=1}^{K} \hat{r}_{jk} \hat{r}_{j'k} = \sum_{k=1}^{K} \hat{c}_{j,j',k} \hat{\alpha}_{j,j'} + \sum_{k=1}^{K} \hat{d}_{j,j',k} \hat{\alpha}_{j,j'}^2
\]

using a second-order approximation. More details on these approximations, including their derivations and accuracy, are available in Appendix I.
3.3.3 Model Fitting Algorithm

We propose the following ES algorithm to obtain parameter estimates when modeling multiple diseases with group testing data:

1) Select initial values $\beta^{(0)}$ of $\beta$.

2) E-Step: For a given $\beta^{(b)}$, $b = 0, 1, 2, \ldots$, calculate $\omega_k^{(b)} = (E(\tilde{Y}_{1ik} \mid z_{1ik}, \beta^{(b)}), \ldots, E(\tilde{Y}_{Jik} \mid z_{Jik}, \beta^{(b)}))'$, $i = 1, \ldots, I_k$ and $k = 1, \ldots, K$.

3) S-Step: Estimate $\alpha$ using the methods in Section 3.3.2 with the current estimate $\beta^{(b)}$, and denote it as $\hat{\alpha}(\beta^{(b)})$. Solve

$$
\Psi(\beta, \hat{\alpha}(\beta^{(b)})) = \sum_k \sum_i D_i \mathbf{V}_ik^{-1}(\omega_k^{(b)} - \hat{p}_ik) = 0
$$

for $\beta$ to update the parameter estimates at the $(b + 1)^{th}$ iteration, where $\hat{\alpha}(\beta^{(b)})$ within $\mathbf{V}_ik$ and $\omega_k^{(b)}$ are treated as fixed and known.

4) Repeat Steps 2 and 3 until $\|\beta^{(b+1)} - \beta^{(b)}\|$ is very small; denote the final solution by $\hat{\beta}$.

3.3.4 Variance Estimation

Elashoff and Ryan (2004) showed that under certain regularity conditions, regression parameter estimators obtained from the ES algorithm are consistent and are asymptotically normal. Consistency and asymptotic normality also hold in our setting but with a small change to the form of $Cov(\hat{\beta})$. Note that for each group $k$, the expectations $E(\tilde{Y}_{1ik} \mid Z_{ik}), \ldots, E(\tilde{Y}_{I_{ik}} \mid Z_{ik})$ are all functions of $Z_{ik}$; thus the $\Psi_{ik}(\beta, \alpha)$ expressions in the same group are dependent. It is therefore necessary to modify the middle part of the sandwich variance estimator in Elashoff and Ryan (2004) to incorporate this within group dependence. Specifically, the covariance matrix of $\hat{\beta}$ is
\[
\text{Cov}(\hat{\beta}) = \left( \sum_{k} \sum_{i} \frac{\partial \Psi_{ik}^{\text{obs}}(\beta, \alpha)}{\partial \beta} \right)^{-1} \left( \sum_{k} \left( \sum_{i} \Psi_{ik}^{\text{obs}}(\beta, \alpha) \right) \left( \sum_{i} \Psi_{ik}^{\text{obs}}(\beta, \alpha) \right)' \right) \times \\
\left( \sum_{k} \sum_{i} \frac{\partial \Psi_{ik}^{\text{obs}}(\beta, \alpha)}{\partial \beta} \right)^{-1},
\]

where \( \alpha, D_{ik}, V_{ik}, \omega_{ik}, \) and \( \hat{p}_{ik} \) are all functions of \( \beta \). An estimate of this covariance matrix, which we denote by \( \hat{\text{Cov}}(\hat{\beta}) \), arises from evaluating Equation (15) at \( \hat{\alpha} \) and \( \hat{\beta} \). Our simulation evidence in Section 3.4 shows that standard errors are well estimated by the corresponding entries in \( \hat{\text{Cov}}(\hat{\beta}) \) and that resulting Wald confidence intervals confer the nominal level in realistic settings.

### 3.4 Simulation Evidence

We have extensively examined via simulation the performance of our proposed methodology in realistic group testing settings. For illustration, consider a logistic regression model for two diseases and two covariates:

\[
\logit(\hat{p}_{jk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}
\]

for \( j = 1, 2 \), where the between-disease correlation is \( \text{Corr}(\hat{Y}_{1ik}, \hat{Y}_{2ik}) = \alpha \). We simulate the first covariate \( x_{1ik} \) from a Uniform(0, 1) distribution and the second covariate \( x_{2ik} \) from a gamma(17, 1.4) distribution. The true regression parameters chosen are \( \beta_{01} = -6 \), \( \beta_{02} = -7 \), \( \beta_{11} = 0 \), \( \beta_{12} = 1 \), \( \beta_{21} = 0.1 \), and \( \beta_{22} = 0.1 \). These covariate and parameter configurations provide a mean prevalence of approximately 3% for the first disease and 2% for the second disease, which are at typical prevalence levels where group testing would be used. In Appendix J, we provide histograms of the true individual probabilities for a simulated data set under this model.
We employ the following strategy to simulate the group responses $Z_{jk}$ for $j = 1, 2$ and $k = 1, \ldots, K$. With the individual probabilities from Equation (16) and a given value of $\alpha$, we use the correlated binary data generation procedure of Emrich and Piedmonte (1991) to simulate the $(\tilde{Y}_{i1k}, \tilde{Y}_{i2k})'$ responses. These responses are then randomly assigned to groups. The true, unobserved group responses $\tilde{Z}_{jk}$ are obtained using $\tilde{Z}_{jk} = 1$ if $\sum_{i=1}^{I_k} \tilde{Y}_{ijk} > 0$ and $\tilde{Z}_{jk} = 0$ if $\sum_{i=1}^{I_k} \tilde{Y}_{ijk} = 0$ for disease $j$ and group $k$. Allowing for testing error, the observed group test responses $Z_{jk}$ are then simulated from the appropriate Bernoulli distribution with success probability $\eta_j = \delta_j = 0.95$ for $j = 1, 2$.

The ES algorithm with a second-order approximation is used to estimate $\alpha$ and Equation (16) for each of $B = 1000$ simulated data set, where we estimate only one parameter, say $\beta_2$, for both $\beta_{21} = \beta_{22}$ because these two parameters are assumed to be equal. This is motivated by our analysis of the NPHL data in Section 3.5, in which the hypothesis of sharing parameters across diseases for a certain covariate (i.e., across the levels of $j$) is not rejected. Table 3.1 gives the parameter estimates averaged over 1,000 simulated data sets for various combinations of $\alpha$, $K$, and $I_k$ (“Mean” row). The use of large samples sizes ($K > 500$) is motivated by our experience with the NPHL (see Section 3.5). As expected, the regression parameter estimates on average approach their corresponding parameters as $K$ increases. We also calculate the standard deviation (SD) for each regression parameter estimate across the 1,000 simulated data sets and compare this to the corresponding averaged estimated standard error (SE) that would be obtained from (15). Also as expected, the SE/SD ratio given in Table 3.1 approaches 1 as $K$ increases, although the SE is slightly
underestimated for smaller $K$. Lastly, in Table 3.1, we give the estimated coverage levels of 95% Wald confidence intervals for each regression parameter. These levels are all between 0.94 and 0.96, which indicate the intervals are performing as expected.

It is often of interest to see how the standard errors of joint modeling of all diseases using the ES algorithm compare to fitting separate group testing models to each disease using the method of Vansteelandt et al. (2000). To explore this, we calculate the relative efficiency as

$$\text{RE}(\hat{\beta}_{j,r}^{\text{ES}} \text{ to } \hat{\beta}_{j,r}^{\text{ML}}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{\text{Var}}(\hat{\beta}_{j,r}^{\text{ML}})}{\hat{\text{Var}}(\hat{\beta}_{j,r}^{\text{ES}})}, \quad (17)$$

where $\hat{\beta}_{j,r}^{\text{ML}}$ and $\hat{\beta}_{j,r}^{\text{ES}}$ denote the $r^{th}$ regression parameter estimate for the $j^{th}$ disease using the Vansteelandt et al. (2000) approach and ES algorithm, respectively. Table 3.2 displays the relative efficiencies for the same simulated data in Table 3.1. Note that we calculate the relative efficiency using $\hat{\text{Var}}(\hat{\beta}_{j,2}^{\text{ES}})$ when $r = 2$ because the single parameter $\beta_2$ replaces $\beta_{21} = \beta_{22}$. For relative efficiencies involving $\hat{\beta}_{j,2}^{\text{ES}}$, dramatic increases in efficiency are seen in Table 3.2 with levels at times greater than 2. In addition, even when parameters are not shared for $r = 1$, we still see valuable gains in efficiency ranging from 1.4% to as high as 17.2%. To compare all regression estimators for each $j$, we also include in Table 3.2 the relative efficiency as in Equation (17), but now involving $\hat{\text{Var}}(\text{logit}(\hat{p}_b))$ where $\hat{p}$ denotes the estimated probability of disease positivity at the mean values of the two covariates in Equation (16). Again, we see the benefits of joint modeling with gains in efficiency ranging from 16.3% to 43.1%.
We have performed a number of additional simulations using different models, different prevalence levels, and different levels of correlation among diseases. Details for these simulations are provided in Appendix J. For example, corresponding to Equation (17), we have also calculated the relative efficiencies where $\beta_{21}$ and $\beta_{22}$ are estimated separately. It is not surprising that the relative efficiencies in this situation are smaller, but they are still as large as 11%.

3.5 Applications

Chlamydia and gonorrhea are the two most prevalent sexually transmitted diseases reported in the United States (Centers for Disease Control and Prevention, 2010). This is true in Nebraska as well, and these diseases even have been characterized as being at epidemic levels in Omaha (Zagurski, 2006). As part of the Centers for Disease Control and Prevention funded Infertility Prevention Project (IPP), the NPHL uses a single assay to test for chlamydia and gonorrhea simultaneously. Due to the high cost incurred by their use of individual testing (approximately $11 for a swab test and $16 for a urine test) and the large numbers of individuals tested (approximately 25,000 per year), the NPHL is interested in using group testing for screening. A few other laboratories, such as the State Hygienic Laboratory at the University of Iowa, already use group testing as part of their participation in the IPP. Our goal is to fit models that can estimate an individual’s probability of having chlamydia or gonorrhea using group testing responses. This would enable our medical colleagues at the NPHL to understand how these disease statuses are related to certain risk factors at a fraction of the cost when compared to testing subjects individually.
Furthermore, the models could also provide additional insight on how to retest individuals in positive groups if identification of positive and negative individuals was our goal (Bilder, Tebbs, and Chen 2010a).

We focus on the 14,530 female swab specimens that were tested individually by the NPHL in 2009. The overall prevalence for chlamydia and gonorrhea during this year was approximately 0.069 and 0.013, respectively (unadjusted for potential testing error). We construct groups of size 5 with the observed data by assigning individuals to groups based on specimen arrival date. Groups of this or of similar size are used elsewhere for chlamydia and gonorrhea screening; see Morre et al. (2001) and Butylkina et al. (2007). The NPHL’s assay for female swabs has a sensitivity of 0.928 for chlamydia and 0.966 for gonorrhea and a specificity of 0.960 for chlamydia and 0.980 for gonorrhea. We use these same levels in our analysis here. In addition to the testing outcomes for both infections, the NPHL collects additional covariate information on each individual. Specifically, we use the following covariates in our models: age, race (represented by three indicator variables), symptoms, clinician observation variables (cervical friability, pelvic inflammatory disease, cervicitis), and risk history variables (multiple partners, new partner in the last 90 days, contact with someone who has a sexually transmitted disease). All covariates are dichotomous (0 and 1) except for age.

Table 3.3 displays the results from fitting a first-order model using our methodology in Section 3.3 with a logit link function. The estimated value of $\alpha$ is 0.27, which is obtained through using a second-order approximation for it (see Section 3.3.2). For comparison purposes, we also fit the same regression model
using the individual observations with standard GEE methodology. This is why we use data that were originally collected on each individual; otherwise, it would not be possible to make this type of comparison. When we fit the individual testing model, we assume that the assay sensitivity and specificity are equal to 1. We attempted to fit this model using the GEE methodology of Neuhaus (2002), which allows for imperfect sensitivity and specificity, but many of the parameter estimates associated with gonorrhea infections did not converge. A further investigation on our part revealed that this is caused by a low gonorrhea prevalence at the given specificity level. In fact, the maximum likelihood estimate for the overall gonorrhea prevalence is actually negative.

The parameter estimates given in Table 3.3 for the group and individual testing models are often in close agreement. The estimated standard errors associated with individual testing are lower than those of the group testing models. This is expected because there are five times more observations used to fit the individual testing model; see Vansteelandt et al. (2000) and Bilder and Tebbs (2009) for a similar discussion with single-disease group testing models. However, it is interesting to note that the group testing standard errors are only 1.3 to 3.2 times more than the individual testing standard errors.

Using a 0.05 level of significance with the group testing models, Wald test p-values (not shown) are less than 0.05 for the covariates:

- Race, symptoms, multiple partners, and contact to a STD corresponding to gonorrhea, and
- Age, race, symptoms, cervicitis, and contact to a STD corresponding to chlamydia.
In this assessment, we perform one test jointly for the four race levels. These results largely agree with those from fitting the individual testing model, although the individual testing analysis finds some additional parameters significant at the 0.05 level (age, pelvic inflammatory disease, and cervicitis for gonorrhea).

Using our group testing model, it is possible to perform hypothesis tests of the form $H_0 : \beta_{r_1} = \beta_{r_2}$ versus $H_a : \beta_{r_1} \neq \beta_{r_2}$ for $r = 0, 1, \ldots, p - 1$; i.e., we can test for a common parameter across diseases. It is important to emphasize that these tests cannot be performed using single-disease group testing regression models, because parameters are estimated separately for each infection. The following covariates have large Wald test $p$-values using the group testing model: pelvic inflammatory disease ($p$-value = 0.642), new partner ($p$-value = 0.533), cervicitis ($p$-value = 0.516), and cervical friability ($p$-value = 0.466). In the light of these findings, it may be reasonable to consider a more parsimonious model with a shared parameter across both diseases for these covariates. When we estimate this model (see Appendix K), we find that Wald test $p$-values are generally less than 0.05 for the same covariates as before. The only difference is that the significant parameter for cervicitis is now shared across the infections.

### 3.6 Discussion

In this chapter, we have developed a group testing regression model based on the ES algorithm for correlated multiple-disease data. Specifically, our proposed method takes advantage of covariate information in estimating individual statuses of multiple diseases simultaneously in a group testing setting. Also, our
methods are especially useful when comparisons of model parameters across diseases are desired. R functions are created for implementing this procedure and are available at www.chrisbilder.com/grouptesting/multiple. In the future, we intend to include the functions in R’s binGroup package (Bilder et al. 2010b; R Development Core Team 2012).

We also derived another approach (called “GEE-group”) for modeling correlated multiple-diseases, where we focus on the observed group responses $Z_k = (Z_{1k}, \ldots, Z_{jk})'$ and solve the estimating equations written in terms of $Z = (Z_1, \ldots, Z_K)'$. This method can be viewed as a direct generalization of the Vansteelandt et al. (2000) approach for single-disease group testing models. GEE-group produces estimates close to those given by the ES algorithm. However, this method has a couple of clear drawbacks that would keep us from using it. First, the working correlation structure must be specified in terms of group responses, which is far less natural than being specified through individual responses like in the ES algorithm. Second, the GEE-group approach cannot be generalized to incorporate any retesting information while we expect the ES algorithm-based method could be generalized (see the last paragraph). More details on our investigation into the GEE-group approach can be found in Section 4.5.

If we let the $j$ subscript in our notation represent time points rather than different diseases, our proposed method is directly applicable to a single-disease longitudinal testing setting. This modeling approach restricts individual subjects to be within the same groups over time. We tried to relax this restriction by allowing individuals to appear in different groups at different time points, but we
found it is mathematically quite difficult because $Z_{jk}$ now could be correlated with $Z_{jk'}$ for any $k'$ (i.e., responses are correlated across groups).

Using only the group responses, our proposed method can estimate covariate-adjusted individual probabilities with reduced cost. When further identification of positive individuals is needed, retesting individuals (or subsets of individuals) from positive groups is often performed. Future research should examine how to incorporate the individual retest outcomes into the estimation process. We expect these additional responses will lead to improved parameter estimates. Our ES algorithm-based approach most likely could accommodate these situations by taking into account the retests in the conditional expectations of the E-step. For some retesting schemes, these conditional expectations may not be available in closed form, but the Gibbs sampling technique could be employed to approximate them. One complication of including the retest results is that for different diseases we will likely have different positive groups, leading to different groups of individuals being retested for $j = 1, 2, \ldots, J$. Consequently, how to effectively make use of different individual subjects in estimating the within-subject correlation is challenging and remains a good future research topic.

3.7 References


Zagurski, K. (2006). Douglas County rates B+ on meeting its health goals, but Dr. Adi Pour says there’s ‘A lot of work to be done’ on reducing STDs. *Omaha World Herald* February 2, p. 08B.
Table 3.1. Simulation results from using the ES algorithm to estimate the model in Equation (16) with $\beta_{01} = -6$, $\beta_{02} = -7$, $\beta_{11} = 0$, $\beta_{12} = 1$, and $\beta_2 = 0.1$. A second-order approximation is used to estimate $\alpha$ as described in Section 3.3.2. Estimated parameters and standard errors are averaged over 1000 simulated data sets. Estimated coverage probabilities are for nominal 95% Wald confidence intervals.

<table>
<thead>
<tr>
<th>$\alpha = 0.6$, $I_k = 5$, $K = 1000$</th>
<th>$\alpha = 0.6$, $I_k = 10$, $K = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>$\beta_{01}$</td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>$\beta_{02}$</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$\beta_{11}$</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$\beta_{12}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>SE/SD</td>
<td>SE/SD</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
</tr>
<tr>
<td>$\alpha = 0.2$, $I_k = 5$, $K = 1000$</td>
<td>$\alpha = 0.2$, $I_k = 10$, $K = 500$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>$\beta_{01}$</td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>$\beta_{02}$</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$\beta_{11}$</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$\beta_{12}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>SE/SD</td>
<td>SE/SD</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
</tr>
<tr>
<td>$\alpha = 0.6$, $I_k = 5$, $K = 2000$</td>
<td>$\alpha = 0.6$, $I_k = 10$, $K = 1000$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>$\beta_{01}$</td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>$\beta_{02}$</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$\beta_{11}$</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$\beta_{12}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>SE/SD</td>
<td>SE/SD</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
</tr>
<tr>
<td>$\alpha = 0.2$, $I_k = 5$, $K = 2000$</td>
<td>$\alpha = 0.2$, $I_k = 10$, $K = 1000$</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>$\beta_{01}$</td>
<td>$\beta_{01}$</td>
</tr>
<tr>
<td>$\beta_{02}$</td>
<td>$\beta_{02}$</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$\beta_{11}$</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$\beta_{12}$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>
Table 3.2. Relative efficiency of the variance estimates for the model in Equation (16). A single parameter $\beta_2$ is estimated for both $\beta_{21} = \beta_{22}$ by the ES algorithm.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$K$</th>
<th>$I_k$</th>
<th>$\hat{\beta}_{01}$</th>
<th>$\hat{\beta}_{02}$</th>
<th>$\hat{\beta}_{11}$</th>
<th>$\hat{\beta}_{12}$</th>
<th>$\hat{\beta}_{21}$</th>
<th>$\hat{\beta}_{22}$</th>
<th>$\text{logit}(\hat{p})$</th>
<th>$j = 1$</th>
<th>$j = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1000</td>
<td>5</td>
<td>1.249</td>
<td>1.669</td>
<td>1.085</td>
<td>1.067</td>
<td>1.290</td>
<td>2.229</td>
<td>1.211</td>
<td>1.279</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>10</td>
<td>1.287</td>
<td>1.737</td>
<td>1.113</td>
<td>1.172</td>
<td>1.335</td>
<td>2.358</td>
<td>1.287</td>
<td>1.431</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>1000</td>
<td>5</td>
<td>1.409</td>
<td>1.828</td>
<td>1.049</td>
<td>1.088</td>
<td>1.573</td>
<td>2.598</td>
<td>1.174</td>
<td>1.326</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>10</td>
<td>1.469</td>
<td>1.897</td>
<td>1.079</td>
<td>1.136</td>
<td>1.718</td>
<td>2.817</td>
<td>1.264</td>
<td>1.404</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>2000</td>
<td>5</td>
<td>1.197</td>
<td>1.575</td>
<td>1.050</td>
<td>1.014</td>
<td>1.237</td>
<td>1.984</td>
<td>1.163</td>
<td>1.224</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.242</td>
<td>1.584</td>
<td>1.061</td>
<td>1.074</td>
<td>1.312</td>
<td>1.999</td>
<td>1.218</td>
<td>1.264</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>2000</td>
<td>5</td>
<td>1.373</td>
<td>1.733</td>
<td>1.016</td>
<td>1.032</td>
<td>1.521</td>
<td>2.411</td>
<td>1.173</td>
<td>1.275</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.462</td>
<td>1.758</td>
<td>1.038</td>
<td>1.070</td>
<td>1.655</td>
<td>2.455</td>
<td>1.241</td>
<td>1.340</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3. Parameter estimates and estimated standard errors (in parentheses) for the NPHL data. The GEE column corresponds to a model fit to the individual testing responses using GEE methodology.

<table>
<thead>
<tr>
<th>Term</th>
<th>Disease</th>
<th>ES algorithm</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>-5.722(0.605)</td>
<td>-4.553(0.327)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>-0.520(0.419)</td>
<td>-0.976(0.147)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>-0.031(0.021)</td>
<td>-0.040(0.013)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>-0.113(0.019)</td>
<td>-0.088(0.007)</td>
<td></td>
</tr>
<tr>
<td>Race level #1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>2.020(0.359)</td>
<td>1.319(0.173)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.591(0.120)</td>
<td>0.392(0.096)</td>
<td></td>
</tr>
<tr>
<td>Race level #2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>0.771(1.080)</td>
<td>0.715(0.336)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>1.062(0.243)</td>
<td>0.691(0.136)</td>
<td></td>
</tr>
<tr>
<td>Race level #3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>0.782(0.857)</td>
<td>-0.113(0.425)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.036(0.401)</td>
<td>0.057(0.151)</td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>1.092(0.384)</td>
<td>0.930(0.175)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.385(0.175)</td>
<td>0.287(0.082)</td>
<td></td>
</tr>
<tr>
<td>Cervical friability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>-0.194(0.648)</td>
<td>0.325(0.312)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.309(0.305)</td>
<td>0.056(0.170)</td>
<td></td>
</tr>
<tr>
<td>Pelvic inflammatory disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>0.283(0.963)</td>
<td>1.158(0.524)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.788(0.627)</td>
<td>0.400(0.387)</td>
<td></td>
</tr>
<tr>
<td>Cervicitis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>0.293(0.349)</td>
<td>0.550(0.200)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.534(0.199)</td>
<td>0.591(0.107)</td>
<td></td>
</tr>
<tr>
<td>Multiple partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>1.167(0.311)</td>
<td>1.046(0.171)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.279(0.221)</td>
<td>0.468(0.100)</td>
<td></td>
</tr>
<tr>
<td>New partner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>0.292(0.332)</td>
<td>-0.086(0.186)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.064(0.197)</td>
<td>-0.044(0.092)</td>
<td></td>
</tr>
<tr>
<td>Contact to a STD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gonorrhea</td>
<td>1.381(0.286)</td>
<td>1.170(0.181)</td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>0.591(0.212)</td>
<td>0.935(0.101)</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4: Additional research

4.1 Introduction

In this chapter, we include additional research performed for this dissertation that did not fit into previous chapters. For single-disease group testing models, we derived the explicit form of the likelihood function for the Dorfman and halving protocols, and evaluated the small-sample performance of the likelihood ratio tests through simulation. We then discuss how to extend the ES algorithm in Chapter 3 to allow for the presence of individual retests that would arise from a group testing protocol as mentioned for Section 2.2. For multiple-disease group testing data, we also consider two alternative approaches other than the ES algorithm to estimate the model. The first approach uses random effects to account for the correlation between disease statuses. The second approach constructs a set of generalized estimating equations in terms of the observed group responses \( Z_k = (Z_{1k}, Z_{2k}, \ldots, Z_{jk})' \), rather than in terms of the unobserved individual responses as shown in Chapter 3. Finally, we provide directions for future research on group testing regression models.

4.2 Likelihood Function for Dorfman and Halving Protocols

This sub-section derives the likelihood function for the Dorfman and halving protocol so that a direct maximization of the function can be carried out. A benefit from direct maximization is that deviance statistics can be formed, which
subsequently leads to the construction of likelihood ratio tests (LRTs) and Akaike’s information criterion (AIC) statistics for model comparisons.

For Dorfman’s protocol, using the same notation as defined in Section 2.2, the joint probability for all observed responses can be written as

\[ P(Y = y, Z_1 = z_1, \ldots, Z_K = z_K) \]

where \( y \) is the vector of individual retests within positive pools. Because groups are independent of each other, the above probability can be further written as

\[
\prod_{k:Z_k=1} P(Y_{1k} = y_{1k}, \ldots, Y_{l_kk} = y_{l_kk}, Z_k = 1) \times \prod_{k:Z_k=0} P(Z_k = 0).
\]

Hence, the log-likelihood function is

\[
\sum_{k:Z_k=0} \log(1 - \theta_k) + \sum_{k:Z_k=1} \log P(Y_{1k} = y_{1k}, \ldots, Y_{l_kk} = y_{l_kk}, Z_k = 1),
\]

where \( \theta_k \) is given by Equation (1). Note that \( P(Y_{1k} = y_{1k}, \ldots, Y_{l_kk} = y_{l_kk}, Z_k = 1) \) is the denominator of Equation (23) in Appendix A, where we showed it can be written as

\[
\eta \prod_{i'=1}^{l_k} \sum_{\bar{y}_{i'k}} P(Y_{i'k} = y_{i'k} | \bar{Y}_{i'k} = \bar{y}_{i'k}) P(Y_{\bar{i'}k} = \bar{y}_{\bar{i'}k}) + \varphi \prod_{i'=1}^{l_k} P(Y_{i'k} = y_{i'k} | \bar{Y}_{i'k} = 0)(1 - \bar{p}_{i'k}).
\]

Substituting Equation (19) and Equation (1) into Equation (18), the log-likelihood function for Dorfman’s protocol can be explicitly expressed. A Newton-Raphson procedure can be employed to maximize the log-likelihood with respect to \( \beta \). The inverse of the observed information matrix can serve as the estimated covariance matrix of \( \hat{\beta} \).

Similarly for the halving protocol, the joint probability for all observed responses is a product of \( P(Z_k = 0) \), \( P(Z_k = 1, Z_{kl} = 0, Z_{l2} = 0) \), \( P(Z_k = 1, Z_{kl} = 1, Y_{kl} = y_{kl}, Z_{l2} = 0) \), \( P(Z_k = 1, Z_{kl} = 0, Z_{l2} = 1, Y_{l2} = y_{l2}) \), and \( P(Z_k = 1, Z_{kl} = 1, Y_{kl} = y_{kl}, Z_{l2} = 0) \).
1, $Y_{i1} = y_{i1}$, $Z_{i2} = 1$, $Y_{i2} = y_{i2}$). Each of the above probabilities can be written as a function of the individual probabilities $\tilde{p}_{ik}$, as shown in Appendix B. We can then maximize the log-likelihood to obtain the MLE.

We have verified that for the same data set, the direct maximization of the log-likelihood gives practically the same estimates to those given by the EM algorithm (any differences are due to the convergence criteria). However, we still prefer to use the EM algorithm to estimate parameters. This is because the log-likelihood (especially for halving) is a very complicated function of $\beta$. When the number of predictors in the model is large, the Newton-Raphson procedure can be very slow. Also, the log-likelihood for array testing protocols does not have a closed form solution, and consequently the EM algorithm is necessary for array testing protocols.

We performed a small simulation study to examine the asymptotic distribution of a LRT statistic. In our study, we simulate 1000 data sets, where each contains 5000 individual responses generated with an overall prevalence $p$. The covariates are simulated from a Gamma(17, 1.4) distribution and the group responses are formed with group size of $I$. Group, subgroup, and individual test responses for Dorfman and halving are simulated using $\eta$ and $\delta$ as Bernoulli success probabilities. For each data set, we fit the group testing model

$$\text{logit}(\tilde{p}_{ik}) = \beta_0 + \beta_1 x_{ik}$$

to the responses, and calculate the LRT statistic for testing $H_0$: $\beta_1 = 0$. A Kolmogorov-Smirnov test is performed on the simulated test statistic values to determine if they follow a $\chi^2(1)$ distribution. The results showed that for different combinations of $I$, $\eta$, $\delta$ and $p$, the $\chi^2(1)$ approximation works well for Dorfman and halving.
4.3 ES Algorithm when Retesting Information is Available

When retests of individuals or groups of individuals are performed for a group testing protocol, we can incorporate the retest results into the estimation process for the ES algorithm developed in Chapter 3. Similar to our work in Section 2.2 for single-disease models, we can reformulate the conditional expectations in Section 3.3.1 by taking into account the specific group testing protocol used. When it is not possible to obtain a closed-form expression for these conditional expectations, one can use Gibbs sampling, as demonstrated in Section 2.2.4, to approximate them.

The expressions for the conditional expectations $\omega_{ijk}$ can be easily found for each group testing protocol by adding the extra subscript $j$ to each term in the expressions derived in Section 2.2. We demonstrate this here for Dorfman’s protocol. If a group tests positive for disease $j$ ($Z_{jk} = 1$), then all individuals within it will be individually retested. Denote these binary retest outcomes as $Y_{ijk}$. We need to find the conditional means of the true individual responses given the group responses. For observed negative groups where $Z_{jk} = 0$ so that no retests are performed, the conditional mean is

$$E(\hat{Y}_{ijk} \mid Z_{jk} = 0) = (1 - \eta_j)\tilde{p}_{ijk} / (1 - \theta_{jk}).$$

For observed positive groups where $Z_{jk} = 1$, the conditional mean $\omega_{ijk}$ can be expressed as

$$\omega_{ijk} = P(\hat{Y}_{ijk} = 1 \mid Y_{i,j,k} = y_{i,j,k}, \ldots, Y_{i,j,k} = y_{i,j,k}, Z_{jk} = 1) = \frac{\tilde{p}_{ijk}\eta_j\lambda_{ijk}^{(1)} \prod_{i' = 1}^{l_i} \sum_{y_{i'a} = 0}^{1} \lambda_{i'kj}^{(y_{i'a}j)} P(Y_{i'jk} = \tilde{y}_{i'jk})}{\phi_j \prod_{i' = 1}^{l_i} \lambda_{i'kj}^{(1)} (1 - \tilde{p}_{i'jk}) + \eta_j \prod_{i' = 1}^{l_i} \sum_{y_{i'a} = 0}^{1} \lambda_{i'kj}^{(y_{i'a}j)} P(Y_{i'jk} = \tilde{y}_{i'jk})},$$

where $\lambda_{ijk}^{(y_{i'a}j)} = P(Y_{ijk} = y_{ijk} \mid \hat{Y}_{ijk} = \tilde{y}_{ijk})$ and $\phi_j = 1 - \eta_j - \delta_j$. 

With potentially multiple responses for each group (e.g., group responses $Z_{jk}$ and individual retests $Y_{ijk}$), it is not clear what residuals represent from a model’s fit. This is important because the residuals are needed to estimate the individual correlations used in the ES Algorithm. If residuals based on the individual retests are available, we can derive a similar relationship to Equation (12) between $\text{Corr}(Y_{ijk}, Y_{ijk})$ and $\text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ijk})$. However, some groups do not have any retesting. Also, individual retests are not always performed for all diseases being tested for a group, so we may not have both $Y_{ijk}$ and $Y_{ij'k}$ for $j \neq j'$. Thus, it is usually not possible to estimate $\text{Corr}(Y_{ijk}, Y_{ij'k})$ by including the retesting information. As a result, we generally can use only the initial group responses to estimate the correlation between diseases statuses. The estimation process is then the same as described in Section 3.3.

4.4 Group Testing Model with Random Effects for Multiple-Disease Data

The inclusion of random effects within a model is a standard technique used to account for within-subject correlations in situations such as longitudinal data analysis. In group testing contexts, Chen et al. (2009) is the only paper that has incorporated random effects into a group testing regression model, and this research was for the single-disease setting only. This sub-section proposes two ways to estimate a model that includes a random effect to account for the within subject correlation that occurs when multiple disease responses are observed. Thus, this model could serve as an alternative to the methods described in Chapter 3. Note the proposals given here have not been implemented due to
expected computational difficulties. We present these two proposals as a record of our research activities, and we hope that these proposals could serve as guidance for future research in the area.

We continue to use the same notation as defined in Section 3.2. Regarding each individual subject’s responses as a cluster, our model has the form:

$$\text{logit}(P(\tilde{Y}_{ijk} = 1 | u_{ik})) = X_{ik} \beta_j + u_{ik}$$

where $X_{ik} = (x_{1ik}, ..., x_{p-1,ik})'$ are the covariates, $\beta_j$ is a $p \times 1$ vector of fixed effects parameters, and $u_{ik}$ are i.i.d. $\sim N(0, \sigma^2)$. We assume that each $\tilde{Y}_{ijk}$ are independent across the subjects $i = 1, ..., I_k$ within each group $k$, and we allow for each $\tilde{Y}_{ijk}$ to be independent for $j = 1, ..., J$ given $u_{ik}$. Note that $\tilde{Y}_{ijk}$ are not observed, and we use only the observable group responses $Z_{jk}$ to obtain parameter estimates (no retests are performed). Within this setting, we can write the probability of a group testing positive as

$$P(Z_{jk} = 1 | u_i) = \eta_j + (1 - \eta_j - \delta_j) \prod_{i=1}^{I_k} (1 - P(\tilde{Y}_{ijk} = 1 | u_{ik}))$$

$$= \eta_j + (1 - \eta_j - \delta_j) \prod_{i=1}^{I_k} \frac{1}{1 + \exp(X_{ik} \beta_j + u_{ik})}.$$

Let $Z_k = (Z_{1k}, Z_{2k}, ..., Z_{Jk})'$ denote a random vector of the trait responses for group $k = 1, ..., K$. We can concatenate these vectors as $Z = (Z_1', Z_2', ..., Z_K')'$ to form a vector of all group responses. We can write the joint density for $Z$ as

$$f(z) = \prod_{k=1}^{K} f(z_k)$$

where $f(\cdot)$ denotes a probability distribution function. Note the equality above is due to the independence of $Z_1, Z_2, ..., Z_K$ (responses are independent across groups). The joint density function for $Z_k$ is

$$f(z_k) = \int f(z_k | u_k) f(u_k) du_k = \int f(z_k | u_k) f(u_k) du_k,$$
where $u_k = (u_{1k}, u_{2k}, \ldots, u_{jk})'$ is the vector of random effects for all individuals in group $k$. In the above equation, $f(u_k) = \prod_{i=1}^{l_k} f(u_{ik})$ due to the independence of random effects across individuals. To find $f(z_k \mid u_k)$ in the above equation, let $\tilde{y}_{jk} = (\tilde{y}_{1jk}, \ldots, \tilde{y}_{ljk})'$ so that we can write the joint density of the individual responses as

$$f(\tilde{y}_{1k}, \ldots, \tilde{y}_{jk} \mid u_k) = \prod_{j=1}^{l_k} f(\tilde{y}_{jk} \mid u_k)$$

for each group $k$. We can express $f(z_k \mid u_k)$ then as

$$f(z_k \mid u_k) = \sum_{\tilde{y}_{1k}, \ldots, \tilde{y}_{jk}} \left\{ f(z_k \mid u_k, \tilde{y}_{1k}, \ldots, \tilde{y}_{jk}) f(\tilde{y}_{1k}, \ldots, \tilde{y}_{jk} \mid u_k) \right\}$$

$$= \sum_{\tilde{y}_{1k}, \ldots, \tilde{y}_{jk}} \left\{ \prod_{j=1}^{l_k} f(z_k \mid \tilde{y}_{jk}) \times \prod_{j=1}^{l_k} f(\tilde{y}_{jk} \mid u_k) \right\}$$

$$= \sum_{\tilde{y}_{1k}, \ldots, \tilde{y}_{jk}} \left\{ \prod_{j=1}^{l_k} f(z_k \mid \tilde{y}_{jk}) \prod_{i=1}^{l_k} f(\tilde{y}_{ik} \mid u_k) \right\}$$

$$= \prod_{j=1}^{l_k} \left\{ \sum_{\tilde{y}_{jk}} f(z_k \mid \tilde{y}_{jk}) \prod_{i=1}^{l_k} f(\tilde{y}_{ik} \mid u_k) \right\}$$

$$= \prod_{j=1}^{l_k} f(z_k \mid u_k).$$

In the above derivation, we make use of the assumption that $f(z_k \mid u_k, \tilde{y}_{1k}, \ldots, \tilde{y}_{jk}) = f(z_k \mid \tilde{y}_{1k}, \ldots, \tilde{y}_{jk}) = \prod_{j=1}^{l_k} f(z_{jk} \mid \tilde{y}_{jk})$. This assumption follows due to constant sensitivity and specificity levels once the true individual responses are known and the Litvak et al. (1994) discussion that the test outcomes are conditionally independent given the true outcomes. Summarizing, we obtain

$$f(z_k) = \int f(z_k \mid u_k) f(u_k) d\mu_k$$

$$= \int_{\mathbb{R}^{l_k}} \prod_{j=1}^{l_k} f(z_{jk} \mid u_k) \left\{ \prod_{i=1}^{l_k} f(u_{ik}) \right\} d\mu_k.$$
The log likelihood function \( l(\beta, \sigma^2 \mid z) \) can be written as
\[
l(\beta, \sigma^2 \mid z) = \sum_{k=1}^{K} \log(f(z_k)) = \sum_{k=1}^{K} \log\left[ \int \prod_{j=1}^{I_k} f(z_{jk} \mid u_k) \prod_{i=1}^{I_k} f(u_{ik}) \, du_k \right]
\]
where \( \beta = (\beta_1', \beta_2', \ldots, \beta_I')' \) is the vector of all fixed effects,
\[
f(z_{jk} \mid u_k) = \left( \eta_j + (1 - \eta_j - \delta_j) \prod_{i=1}^{I_k} \frac{1}{1 + \exp(X_{ik} \beta_j + u_{ik})} \right)^{z_{jk}} \times \left( 1 - \eta_j - (1 - \eta_j - \delta_j) \prod_{i=1}^{I_k} \frac{1}{1 + \exp(X_{ik} \beta_j + u_{ik})} \right)^{1-z_{jk}}
\]
and
\[
f(u_{ik}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u_{ik}^2}{2\sigma^2}\right).
\]
To maximize the log-likelihood function and find the corresponding maximum likelihood estimates (MLEs), one could approximate the integral in Equation (20) by using adaptive Gauss-Hermite quadrature (Pinheiro and Bates 1995). We expect this method to work when the group size \( I_k \) is small, but it would become computationally expensive when the group size gets large. This is because the likelihood function \( l(\beta, \sigma^2 \mid z) \) involves \( K \) different \( I_k \) dimensional integrals. Therefore, when the group size \( I_k \) is large, evaluating the likelihood function directly may be difficult or even intractable.

An alternative way to find maximize the likelihood function and find the corresponding MLEs is through a modified version of the Monte Carlo expected maximization (MCEM) algorithm described in Chen et al. (2009). First, the complete joint log-likelihood function can be expressed as
\[ l(\beta, \sigma^2 \mid z, u) = \log(f(z \mid u)f(u)) \]
\[ = \log f(z \mid u) + \log f(u) \]
\[ = \sum_{k=1}^{K} \sum_{j=1}^{I} \log f(z_{jk} \mid u_k) + \sum_{k=1}^{K} \sum_{i=1}^{b_k} \log f(u_{ik}) \]
\[ = \sum_{k=1}^{K} \sum_{j=1}^{I} \log f(z_{jk} \mid u_k) - \sum_{k=1}^{K} \sum_{i=1}^{b_k} \left[ \log(\sqrt{2\pi}\sigma) + \frac{u_{ik}^2}{2\sigma^2} \right] \]
\[ = \sum_{k=1}^{K} \sum_{j=1}^{I} \log f(z_{jk} \mid u_k) - \left\{ N \log(\sqrt{2\pi}\sigma) + \sum_{k=1}^{K} \frac{u_{ik}^2}{2\sigma^2} \right\}, \]

where \( N = \sum_{k=1}^{K} I_k \) is the total number of individuals and we treat \( u_{ik} \) as missing.

To simplify the notation, define

\[ I_1 = \sum_{k=1}^{K} \sum_{j=1}^{I} \log f(z_{jk} \mid u_k; \beta) \]
\[ I_2 = -N \log(\sqrt{2\pi}\sigma) - \sum_{k=1}^{K} \sum_{i=1}^{b_k} \frac{u_{ik}^2}{2\sigma^2}, \]

where we include \( \beta \) in \( f(z_{jk} \mid u_k; \beta) \) now to emphasize the fixed effect only appears in \( I_1 \). Also, notice that the variance component \( \sigma \) appears only in \( I_2 \). Therefore, we can maximize both parts separately in the M-step to obtain \( \hat{\beta} \) and \( \hat{\sigma}^2 \). Given an initial estimate of the parameters, say, \((\beta^{(b)}, \sigma^{(b)})\), we could use the Metropolis-Hastings algorithm to estimate \( E(I_1|z) \) and \( E(I_2|z) \) in the E-step because \( f(u_k \mid z_i; \beta^{(b)}, \sigma^{(b)}) \) can not be expressed in a closed form. The algorithm generates a large number of samples from \( f(u_k \mid z_i; \beta^{(b)}, \sigma^{(b)}) \), and then use the sample means to estimate \( E(I_1|z) \) and \( E(I_2|z) \). The MCEM algorithm is formally given here:

1. Choose starting values \( \beta^{(0)}, \sigma^{(0)} \) of \( \beta, \sigma \).

2. (E-step). For a given \( \beta^{(b)}, \sigma^{(b)}, b = 0, 1, 2, \ldots \), approximate \( E(I_1|z) \) and \( E(I_2|z) \) by
\[ \hat{I}_1^{(b)} = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{j=1}^{J} \log f(z_{jk}, u_{k}^{(m)}; \beta) \]
\[ \hat{I}_2^{(b)} = -N \log(\sqrt{2\pi\sigma}) - \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \frac{(u_{ik}^{(m)})^2}{2\sigma^2}, \]
respectively, where \( u_{k}^{(m)}, m = 1, ..., M, \) are \( M \) draws from the conditional distribution \( f(u_k | z_k; \beta^{(b)}, \sigma^{(b)}) \), \( k = 1, ..., K \), using the Metropolis-Hastings algorithm.

3) (M-step). Update the parameter estimates to the \((b + 1)\)th iteration by
maximizing \( \hat{I}_1^{(b)} \) with respect to \( \beta \) and maximizing \( \hat{I}_2^{(b)} \) with respect to \( \sigma \).

4) Repeat steps 2 and 3 until \( \|\beta^{(b+1)} - \beta^{(b)}\| \) and \( \|\sigma^{(b+1)} - \sigma^{(b)}\| \) are very small.

As Chen et al. (2009) pointed out, the MCEM algorithm is computationally intensive, but is more flexible and can allow for other random effect distributions and different pooling strategies.

### 4.5 GEE-group Approach for Multiple-Disease Data

The purpose of this sub-section is to illustrate how to formulate generalized estimating equations in terms of the observed group responses so that standard GEE methodology can be adapted to group testing problems. This would be analogous to the approach taken by Vansteelandt et al. (2000) for single-disease group testing models, and is an alternative to the ES algorithm fitting approach to account for the correlation among disease responses.

To account for the correlation among different traits within each group, let \( R(\alpha) \) be the working correlation for \( Z_k = (Z_{1k}, Z_{2k}, ..., Z_{Jk})' \), where the matrix depends on a vector of parameters \( \alpha \). The working covariance matrix of \( Z_k \) is then \( V_k = B_k^{1/2} R(\alpha) B_k^{1/2} \), where \( B_k \) is a \( J \times J \) diagonal matrix with diagonal elements \( \theta_{jk}(1 - \theta_{jk}) \). The GEE for the multiple-disease group testing model is
\[
\sum_{k=1}^{K} D_k V_k^{-1}(z_k - \theta_k) = 0, \tag{21}
\]

where \( D_k = \partial \theta_k / \partial \beta \), \( \theta_k = (\theta_{1k}, \ldots, \theta_{Jk})' \), and \( \mathbf{0} \) is a \( pJ \times 1 \) vector of 0’s. Equation (21) differs from the GEE as defined in Liang and Zeger (1986) by only the form of \( D_k \), which is more complicated now due to the relationship between \( \theta_{jk} \) and \( \tilde{p}_{jk} \) given in Equation (8).

To solve the estimating equations in (21) for \( \beta \), we need to estimate \( \alpha \) first. The Pearson residuals of the group responses are given as \( \hat{\rho}_{jk} = (z_{jk} - \hat{\theta}_{jk}) / \sqrt{\hat{\theta}_{jk}(1 - \hat{\theta}_{jk})} \), where \( \hat{\theta}_{jk} \) is the model’s estimate of \( \theta_{jk} \). We can calculate \( \hat{\alpha} \) using these Pearson residuals in the moment estimators proposed by Liang and Zeger (1986). For example, assuming the exchangeable correlation structure where \( \text{Corr}(Z_{jk}, Z_{jk}) = \alpha \) for all \( j \neq j' \), we have

\[
\hat{\alpha} = \frac{1}{\sum_{k=1}^{K} \sum_{j > j'} \hat{\rho}_{jk} \hat{\rho}_{j'k}} \left[ J(J-1)K - Jp \right].
\]

Parameter estimation and large sample normality of the estimators follow from Liang and Zeger (1986). Parameter estimates can be found by iterating between a modified Fisher scoring algorithm for \( \beta \) and estimating \( \alpha \) based on the current estimates of \( \beta \). Large sample properties then follow with \( \hat{\beta} \) having a large sample normal distribution with mean \( \beta \) and covariance matrix

\[
\left\{ \sum_{k=1}^{K} D_k V_k^{-1} D_k \right\}^{-1} \left\{ \sum_{k=1}^{K} D_k V_k^{-1} \text{Cov}(Z_k) V_k^{-1} D_k \right\} \left\{ \sum_{k=1}^{K} D_k V_k^{-1} D_k \right\}^{-1}.
\]

Replacing \( \beta \) with \( \hat{\beta} \), \( \alpha \) with \( \hat{\alpha} \), and \( \text{Cov}(Z_k) \) with \( (z_k - \hat{\theta}_k)(z_k - \hat{\theta}_k)' \) in the above formula gives us the estimated covariance matrix of \( \hat{\beta} \).

Table 4.1 provides the GEE-group simulation results corresponding to the first set of simulations given in Section 3.4. The GEE-group and ES algorithm do provide similar results, despite the potential problems with the GEE-group
approach outlined in Section 3.6. To further explore these potential problems, we calculated $\text{Corr}(Z_{1k}, Z_{2k})$ for each group using Theorem 1 and a select subset of our simulations. From Figure 4.1, we see that the correlation between $Z_{1k}$ and $Z_{2k}$ varies somewhat over the groups; however, it appears that this variation is not large enough to have a substantial effect on performance of the GEE-group method.

We find it interesting that the GEE-group method can attain similar results as the approach using the ES algorithm. This could occur because of the well-known robustness properties of GEE in general. This finding does not void the merit of our ES algorithm approach because 1) the ES algorithm allows one to specify a working correlation structure on the individual scale, and 2) a direct generalization of the ES algorithm that allows one to incorporate retesting information, as described in Section 4.3, is possible. It is also worth noting that our ES algorithm formulation for group testing data may be applicable in other contexts involving latent correlated binary response data.

4.6 Future Research for Group Testing Regression Models

Section 4.2 provides the likelihood function for the Dorfman and halving protocol, and preliminary simulation results suggest that the likelihood ratio tests for the model parameters follow a chi-square distribution with degree of freedom 1. With the likelihood function available, the residual deviance can be easily obtained, but the degree of freedom associated with it is unknown because each individual is observed multiple times (in a group and by itself) with these protocols. If the degrees of freedom for the residual deviance can be determined, the deviance can
serve as a goodness-of-fit (GOF) statistic for the model. Further, it would be interesting to investigate how to generalize the GOF tests proposed by Chen et al. (2009) to incorporate retest results from Dorfman, halving, and other group testing protocols. We expect that the test statistics for the GOF tests in Chen et al. (2009) could be modified accordingly. However, because the observed responses will no longer be independent from each other, the asymptotic distributions of the test statistics could be challenging to obtain.

In addition to Wald tests and likelihood ratio tests, score tests may also be developed for group testing regression models. The EM algorithm theory implies that (e.g., see Heyde and Morton (1996))

$$\frac{\partial \log L(\theta; y)}{\partial \theta} = E\left[ \frac{\partial \log L(\theta; x)}{\partial \theta} \bigg| y \right],$$

where $x$ denotes the complete data, $y$ denotes the observed incomplete data, $L$ is the likelihood function based on either $x$ or $y$. Note that the right hand side of Equation (22) is the conditional score function, which is easily obtainable from the M-step of the EM algorithm. The Louis’s (1982) method gives $-E[\partial^2 \log L(\theta; y) / \partial \theta^2]$. As a result, the score test can be easily constructed from the EM algorithm, and it is a very natural way of testing the parameters for single-disease group testing models. Moreover, it is readily applicable to group testing protocols whose likelihood function can not be explicitly expressed (e.g., array testing). Future work should examine the finite-sample performance of these tests and compare the score tests to Wald and likelihood ratio tests.

Delaigle and Meister (2011) and Delaigle and Hall (2012) proposed a local polynomial regression model for group testing data. They mainly illustrated their
method with a single-covariate model. In the last section, the authors briefly discussed how the kernel-based estimator can be generalized to the multivariate setting. However, in a standard regression context, local regression models are often less useful in higher dimensions (>2), unless we are willing to make simple structure assumptions (e.g., additive models). This is because multidimensional kernel estimators often require burdensome computations. We believe this is also the case for group testing data. In group testing applications, there are often many potential covariates for each individual subject, so the use of their method is somewhat limited in practice. On the other hand, regression splines can be easily extended to non-additive models. In particular, multivariate adaptive regression splines (MARS) is a popular non-parametric regression technique for modeling of high dimensional data. In the future, it would be of great practical interest to investigate how to apply MARS to the group testing setting.

We have also briefly explored a Bayesian approach for group testing regression models. The advantage of the Bayesian approach is that due to the use of Markov Chain Monte Carlo (MCMC) methods, no complex algorithm is needed for parameter estimation as long as enough MCMC samples are generated and the model is correctly specified. This approach can be implemented directly in standard statistical software (e.g., WinBUGS, R2WinBUGS package in R). For example, we consider the following model for multiple-disease data (notation follow from Section 4.4):

\[ \logit(P(Y_{ijk} = 1 \mid u_{ijk}) = X_{ik} \beta_j + u_{ijk}, \]

where
\[ P(Z_{jk} = 1 \mid \mathbf{u}_k) = \eta_j + (1 - \eta_j - \delta_j) \prod_{i=1}^{k_i} (1 - P(\tilde{Y}_{ijk} = 1 \mid u_{ijk})) \]

\[ = \eta_j + (1 - \eta_j - \delta_j) \prod_{i=1}^{k_i} \frac{1}{1 + \exp(\mathbf{X}_{ik}\mathbf{\beta}_j + u_{ijk})}, \]

\[ \mathbf{u}_k = (u_{i1k},...,u_{iik})', \text{ and } \mathbf{u}_k \text{ i.i.d. } \sim N_j(\mathbf{0},\Sigma), \text{ for } i = 1, ..., I_k, \text{ } k = 1, ..., K. \]

The model here is a little different from the one in Section 4.4. More parameters are introduced to allow for a more flexible correlation structure of the disease statuses. We could use non-informative priors on the parameters:

\[ \mathbf{\beta}_j \text{ iid } \sim N_p(\mathbf{0},1000I_j) \text{ for } j = 1, ..., J, \text{ and } \Sigma^{-1} \sim W_j(2,I_j) \text{ where } W \text{ is the Wishart distribution.} \]

A simple simulated data set was fit by the above model in WinBUGS, and estimated posterior densities for \( \mathbf{\beta}_j \) and \( \Sigma \) were obtained. As mentioned earlier, the Bayesian approach does not require a complex algorithm for parameter estimation, and thus is highly flexible and suited for group testing regression models. We believe it is worthwhile to explore this approach extensively for various group testing protocols in the future.
Table 4.1. GEE-group simulation results corresponding to the model in Equation (16) and to the simulations in Table 3.1. The true parameters are $\beta_{01} = -6$, $\beta_{11} = 0$, $\beta_{02} = -7$, $\beta_{12} = 1$, and $\beta_2 = 0.1$. Estimated coverage is given for 95% Wald confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.6$, $I = 5$, $K = 1000$</th>
<th>$\alpha = 0.6$, $I = 10$, $K = 500$</th>
<th>$\alpha = 0.2$, $I = 5$, $K = 1000$</th>
<th>$\alpha = 0.2$, $I = 10$, $K = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
</tr>
<tr>
<td>Mean</td>
<td>-5.98 -7.01 -0.03 0.99 0.10</td>
<td>-6.13 -7.20 0.00 1.08 0.10</td>
<td>-6.02 -7.03 0.03 1.02 0.10</td>
<td>-6.12 -7.22 0.02 1.14 0.10</td>
</tr>
<tr>
<td>SE/SD</td>
<td>0.96 0.96 0.99 0.95 0.94</td>
<td>0.99 0.96 0.93 0.91 0.94</td>
<td>0.95 0.95 0.95 0.95 0.94</td>
<td>0.94 0.95 0.95 0.95 0.95</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.95 0.94 0.96 0.94 0.94</td>
<td>0.95 0.94 0.95 0.95 0.94</td>
<td>0.94 0.95 0.95 0.94 0.94</td>
<td>0.94 0.95 0.95 0.95 0.95</td>
</tr>
<tr>
<td></td>
<td>$\alpha = 0.6$, $I = 5$, $K = 2000$</td>
<td>$\alpha = 0.6$, $I = 10$, $K = 1000$</td>
<td>$\alpha = 0.2$, $I = 5$, $K = 2000$</td>
<td>$\alpha = 0.2$, $I = 10$, $K = 1000$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
<td>$\beta_{01}$ $\beta_{02}$ $\beta_{11}$ $\beta_{12}$ $\beta_2$</td>
</tr>
<tr>
<td>Mean</td>
<td>-6.01 -7.02 0.00 1.02 0.10</td>
<td>-6.01 -7.05 0.04 1.07 0.10</td>
<td>-6.02 -7.05 0.01 1.05 0.10</td>
<td>-6.05 -7.07 0.03 1.03 0.10</td>
</tr>
<tr>
<td>SE/SD</td>
<td>0.99 0.99 0.97 0.98 1.00</td>
<td>0.99 0.99 0.95 0.96 0.99</td>
<td>0.97 0.96 1.01 0.99 0.96</td>
<td>0.97 1.00 0.95 0.99 0.97</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.95 0.94 0.94 0.95 0.96</td>
<td>0.94 0.96 0.94 0.95 0.94</td>
<td>0.94 0.96 0.94 0.95 0.94</td>
<td>0.94 0.96 0.94 0.95 0.94</td>
</tr>
</tbody>
</table>
Figure 4.1. True correlation between disease responses plotted over the groups. The left-side plot is for the first generated data set from the $\alpha = 0.6$, $I = 5$, and $K = 1000$ simulations. The right-side plot is for the first generated data set from the $\alpha = 0.6$, $I = 10$, and $K = 1000$ simulations. The horizontal line is the estimated working correlation found for that data set by using the GEE-group method. Note that other simulated data sets gave very similar results.
References


Centers for Disease Control and Prevention (2012). *Infertility Prevention Project*


Zagurski, K. (2006). Douglas County rates B+ on meeting its health goals, but Dr. Adi Pour says there’s ‘A lot of work to be done’ on reducing STDs. *Omaha World Herald* February 2, p. 08B.
Appendix A

This appendix shows how to find $\omega_{ik}$ for Dorfman’s protocol when $Z_k = 1$. We first express the conditional mean $\omega_{ik}$ as

$$P(\tilde{Y}_{ik} = 1 \mid Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1) = \frac{P(Y_{ik} = 1, Y_{i_{1:k}} = y_{i_{1:k}}, Z_k = 1)}{P(Y_{ik} = 1, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)}$$

$$= \frac{\sum_{\tilde{y}_{-i,k}} P(\tilde{Y}_{ik} = 1, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}, Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)}{\sum_{\tilde{y}_k} P(\tilde{Y}_k = \tilde{y}_k, Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)}, \quad (23)$$

where $\tilde{Y}_k = (\tilde{Y}_{ik}, \ldots, \tilde{Y}_{l_k,k})'$ and $\tilde{Y}_{-i,k}$ is the same as $\tilde{Y}_k$ but without $\tilde{Y}_{i,k}$.

Examining the numerator in (23) before the summation symbol, we have

$$P(\tilde{Y}_{ik} = 1, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}, Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)$$

$$= P(Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)P(\tilde{Y}_{ik} = 1)P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k})$$

$$= \tilde{p}_k P(Y_{ik} = y_{ik} | \tilde{Y}_{ik} = 1) \prod_{i' = 1}^k P(Y_{i',k} = y_{i',k} | \tilde{Y}_{i',k} = \tilde{y}_{i',k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \times$$

$$= \eta \tilde{p}_k P(Y_{ik} = y_{ik} | \tilde{Y}_{ik} = 1) \prod_{i' = 1}^k P(Y_{i',k} = y_{i',k} | \tilde{Y}_{i',k} = \tilde{y}_{i',k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}), \quad (24)$$

where we use the standard assumption that test outcomes are conditionally independent given the true outcomes (Litvak et al. 1994). Similarly, from the denominator of (23), we have

$$P(\tilde{Y}_k = \tilde{y}_k, Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)$$

$$= P(Y_{ik} = y_{ik}, \ldots, Y_{i_{l_k},k} = y_{i_{l_k},k}, Z_k = 1)P(\tilde{Y}_k = \tilde{y}_k)$$

$$= P(Z_k = 1 | \tilde{Y}_k = \tilde{y}_k) \prod_{i' = 1}^k P(Y_{i',k} = y_{i',k} | \tilde{Y}_{i',k} = \tilde{y}_{i',k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}). \quad (25)$$

Substituting (24) and (25) into (23) results in

$$w_{ik} = \frac{\eta \tilde{p}_k P(Y_{ik} = y_{ik} | \tilde{Y}_{ik} = 1) \sum_{\tilde{y}_{-i,k}} \prod_{i' = 1}^k P(Y_{i',k} = y_{i',k} | \tilde{Y}_{i',k} = \tilde{y}_{i',k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik})}{\sum_{\tilde{y}_k} P(Z_k = 1 | \tilde{Y}_k = \tilde{y}_k) \prod_{i' = 1}^k P(Y_{i',k} = y_{i',k} | \tilde{Y}_{i',k} = \tilde{y}_{i',k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik})}. \quad (26)$$
Equation (26) is very difficult to compute for large group sizes, because the number of summands within it increases exponentially as the group size increases (e.g., there are $2^k$ terms to sum in the denominator for group $k$). Fortunately, we can reformulate the numerator and denominator to make Equation (26) computationally feasible for large group sizes. The denominator can be written as

$$
\sum_{y_k} P(Z_k = 1 \mid \tilde{Y}_k = \tilde{y}_k) \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k})
$$

$$
= P(Z_k = 1 \mid \tilde{Z}_k = 0) \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) (1 - \tilde{p}_{i,k}) +
$$

$$
P(Z_k = 1 \mid \tilde{Z}_k = 1) \sum_{y_k} \Pi_{i=0}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k})
$$

$$
= (1 - \delta) \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) (1 - \tilde{p}_{i,k}) +
$$

$$
\eta \left[ \Pi_{i=1}^k \sum_{y_{i,k}} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) - \left( \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) (1 - \tilde{p}_{i,k}) \right) \right]
$$

$$
= \eta \Pi_{i=1}^k \sum_{y_{i,k}} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) +
$$

$$
\varphi \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) (1 - \tilde{p}_{i,k}),
$$

where $\varphi = 1 - \eta - \delta$ and we make use of the relation

$$
\sum_{y_k} \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k})
$$

$$
= \Pi_{i=1}^k \sum_{y_{i,k}} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}).
$$

Using the same technique, the numerator of (26) can be re-written in a similar manner leading to

$$
w_{i,k} = \eta \tilde{p}_{i,k} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 1) \Pi_{i=1}^k \sum_{y_{i,k}} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k})
$$

$$
= \frac{\varphi \Pi_{i=1}^k P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{p}_{i,k}) +
$$

$$
\eta \Pi_{i=1}^k \sum_{y_{i,k}} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \right) \right].
$$

(27)
The denominator (numerator) of Equation (27) is the product of $I_k (I_k - 1)$ terms, rather than the sum of $2^{I_k}$ terms. Therefore, this formula makes finding $\omega_{ik}$ possible even for large group sizes.
Appendix B

This appendix is a web appendix for the Chapter 2 paper submission. We show here how to find $\omega_{ik}$ for scenarios 1) to 5) of the halving protocol. The derivation is very similar to that of Dorfman’s protocol.

1) $Z_k = 0$:
This is the same as given in Section 2.2 for Dorfman’s protocol.

2) $Z_k = 1$, $Z_{k1} = 0$, $Z_{k2} = 0$:
We need to find

$$
\omega_{ik} = P(\tilde{Y}_{ik} = 1 \mid Z_k = 1, Z_{k1} = 0, Z_{k2} = 0) = \frac{P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 0, \tilde{Y}_{ik} = 1)}{P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 0)}.
$$

For the denominator, we begin by including $\tilde{Y}_k = (\tilde{Y}_{i1}, ..., \tilde{Y}_{i|k|})'$ in the expression to obtain

$$
P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 0) = \sum_{\tilde{y}_k} P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 0 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2})P(\tilde{Y}_{k1} = \tilde{y}_{k1})P(\tilde{Y}_{k2} = \tilde{y}_{k2})
$$

$$
= \sum_{\tilde{y}_{k1}} \sum_{\tilde{y}_{k2}} P(Z_k = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2})P(Z_{k1} = 0 \mid \tilde{Y}_{k1} = \tilde{y}_{k1})P(\tilde{Y}_{k1} = \tilde{y}_{k1})
$$

$$
\times P(Z_{k2} = 0 \mid \tilde{Y}_{k2} = \tilde{y}_{k2})P(\tilde{Y}_{k2} = \tilde{y}_{k2})
$$

$$
= \delta^2(1 - \delta) \prod_{i=1}^{k1} (1 - \tilde{p}_{i/k}) + \delta(1 - \eta) \eta \prod_{i \in k1} (1 - \tilde{p}_{i/k}) \left[1 - \prod_{i' \in k2} (1 - \tilde{p}_{i'/k})\right] +
$$

$$
\delta \eta (1 - \eta) \left\{1 - \prod_{i' \in k1} (1 - \tilde{p}_{i'/k})\right\} \prod_{i' \in k2} (1 - \tilde{p}_{i'/k})
$$

$$
+ \eta (1 - \eta)^2 \left\{1 - \prod_{i' \in k1} (1 - \tilde{p}_{i'/k})\right\} \left\{1 - \prod_{i' \in k2} (1 - \tilde{p}_{i'/k})\right\},
$$

where we let $i \in kj$ denote those individuals within the $j$th subgroup ($j = 1, 2$) and we again use the standard assumption that test outcomes are conditionally independent given the true outcomes (Litvak et al. 1994).
Without loss of generality, we assume here and throughout this appendix that individual $i$ is within the first subgroup ($i \in k_1$). The numerator can be written as

$$P(\tilde{Y}_i = 1, Z_k = 1, Z_{k1} = 0, Z_{k2} = 0)$$

$$= \sum_{\bar{y}_{-i,k}} P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 0 \mid \tilde{Y}_i = 1, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k})$$

$$= \sum_{\bar{y}_{-i,k}} \eta(1 - \eta) P(Z_{k2} = 0 \mid \tilde{Y}_{ik} = 1, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \tilde{p}_{ik}$$

$$= \eta(1 - \eta) \tilde{p}_{ik} \sum_{\bar{y}_{-i,k}} P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \sum_{\bar{y}_{k2}} P(Z_{k2} = 0 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{k2} = \tilde{y}_{k2})$$

$$= \eta(1 - \eta) \tilde{p}_{ik} \sum_{\bar{y}_{-i,k}} P(Z_{k2} = 0 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{k2} = \tilde{y}_{k2})$$

$$= \eta(1 - \eta) \tilde{p}_{ik} \delta \prod_{i' \in k2} (1 - \tilde{p}_{i'k}) + (1 - \eta) \left[ 1 - \prod_{i' \in k2} (1 - \tilde{p}_{i'k}) \right],$$

where $\tilde{Y}_{-i,k} = \{\tilde{Y}_{i',k} : i' = 1, ..., k_2, i' \neq i\}$ is the vector of all true individual statuses excluding the $i^{th}$ subject in group $k$ and $\tilde{Y}_{-i,k1} = \{\tilde{Y}_{i',k1} : i' \in k_1, i' \neq i\}$ is the vector of all true individual statuses excluding the $i^{th}$ subject in subgroup $k_1$.

3) $Z_k = 1, Z_{k1} = 1, Z_{k2} = 0$:

We need to find

$$\omega_{ik} = P(\tilde{Y}_i = 1 \mid Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 0)$$

$$= \frac{P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 0, \tilde{Y}_i = 1)}{P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 0)}.$$

Using the same technique as above, we calculate the denominator to be

$$P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 0)$$

$$= \sum_{\bar{y}_k} P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 0 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2}) \times$$

$$P(\tilde{Y}_{k1} = \tilde{y}_{k1}) P(\tilde{Y}_{k2} = \tilde{y}_{k2})$$

$$= \sum_{\bar{y}_{k2}} \sum_{\bar{y}_{k1}} P(Z_k = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2}) P(Z_{k1} = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}) \times$$

$$\prod_{i' \in k1} P(Y_{i'k} = y_{i'k} \mid \tilde{Y}_{i'k} = \tilde{y}_{i'k}) P(\tilde{Y}_{k1} = \tilde{y}_{k1}) P(\tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{k2} = \tilde{y}_{k2})$$
\[
\begin{aligned}
&= \delta(1 - \delta)^2 \prod_{i \in k_1} \left( P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{p}_{i,k}) \right) \prod_{i \in k_2} (1 - \tilde{p}_{i,k}) + \\
&\quad (1 - \delta)\eta(1 - \eta) \prod_{i \in k_1} \left( P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{p}_{i,k}) \right) \left[ 1 - \prod_{i \in k_2} (1 - \tilde{p}_{i,k}) \right] + \\
&\quad \delta\eta^2 \sum_{y_{i,k} = 0} \prod_{i \in k_1} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \prod_{i \in k_2} (1 - \tilde{p}_{i,k}) + \\
&\eta^2 (1 - \eta) \sum_{y_{i,k} = 0} \prod_{i \in k_1} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \left[ 1 - \prod_{i \in k_2} (1 - \tilde{p}_{i,k}) \right].
\end{aligned}
\]

The numerator can be expressed as

\[
P(\tilde{Y}_{k} = 1, Z_k = 1, Z_{k_1} = 1, Y_{k_1} = y_{k_1}, Z_{k_2} = 0) = \sum_{y_{-i,k}} P(Z_k = 1, Z_{k_1} = 1, Y_{k_1} = y_{k_1}, Z_{k_2} = 0 \mid \tilde{Y}_{i,k} = 1, \tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \times P(\tilde{Y}_{-i,k} = 1) \times \]

\[
P(\tilde{Y}_{k} = 1, Z_{k_2} = 0, \tilde{y}_{k_2}) P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \tilde{p}_{ik} = \eta^2 \tilde{p}_{ik} P(\tilde{Y}_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \times \]

\[
\sum_{y_{-i,k}} \left\{ \prod_{i' \in k_1, i' \neq i} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(Z_{k_2} = 0 \mid \tilde{Y}_{k_2} = \tilde{y}_{k_2}) P(\tilde{Y}_{-i,k} = \tilde{y}_{-i,k}) \right\} \times \]

\[
\sum_{y_{k_2}} \left( P(Z_{k_2} = 0 \mid \tilde{Y}_{k_2} = \tilde{y}_{k_2}) P(\tilde{Y}_{k_2} = \tilde{y}_{k_2}) \right) = \eta^2 \tilde{p}_{ik} P(\tilde{Y}_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \times \]

\[
\sum_{y_{-i,k}} \left\{ \prod_{i' \in k_1, i' \neq i} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \times \]

\[
\left\{ \delta \prod_{i' \in k_2} (1 - \tilde{p}_{i'k}) + (1 - \eta) \left[ 1 - \prod_{i' \in k_2} (1 - \tilde{p}_{i'k}) \right] \right\}.
\]

4) \( Z_k = 1, Z_{k_1} = 0, Z_{k_2} = 1 \):

We need to find

\[
\omega_{ik} = P(\tilde{Y}_{ik} = 1 \mid Z_k = 1, Z_{k_1} = 0, Z_{k_2} = 1, Y_{k_2} = y_{k_2}) = \frac{P(Z_k = 1, Z_{k_1} = 0, Z_{k_2} = 1, Y_{k_2} = y_{k_2}, \tilde{Y}_{ik} = 1)}{P(Z_k = 1, Z_{k_1} = 0, Z_{k_2} = 1, Y_{k_2} = y_{k_2})}.
\]

The denominator follows immediately from previous result by interchanging \( k_1 \) and \( k_2 \):
\[ P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 1, Y_{k2} = y_{k2}) \]
\[ = \delta(1 - \delta)^2 \prod_{i \in k2} \left( P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{P}_{i,k}) \right) \prod_{i \in k1} (1 - \tilde{P}_{i,k}) + \]
\[ (1 - \delta)\eta(1 - \eta) \prod_{i \in k2} \left( P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{P}_{i,k}) \right) \left[ 1 - \prod_{i \in k1} (1 - \tilde{P}_{i,k}) \right] + \]
\[ \delta\eta^2 \sum_{\tilde{y}_{i,k} = 0} \left( \prod_{i' \in k2} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) \right) \prod_{i \in k1} (1 - \tilde{P}_{i,k}) + \]
\[ \eta^2(1 - \eta) \sum_{\tilde{y}_{i,k} = 0} \left( \prod_{i' \in k2} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) \right) \prod_{i \in k1} (1 - \tilde{P}_{i,k}) \right]. \]

We can show the numerator has the following form:
\[ P(\tilde{Y}_{ik} = 1, Z_k = 1, Z_{k1} = 0, Z_{k2} = 1, Y_{k2} = y_{k2}) \]
\[ = \sum_{\tilde{y}_{i,k}} P(Z_k = 1, Z_{k1} = 0, Z_{k2} = 1, Y_{k2} = y_{k2} \mid \tilde{Y}_{i,k} = 1, \tilde{Y}_{i,k} = \tilde{y}_{i,k}) \times \]
\[ P(\tilde{Y}_{ik} = 1, \tilde{Y}_{i,k} = \tilde{y}_{i,k}) \]
\[ = \sum_{\tilde{y}_{i,k}} \eta(1 - \eta) \prod_{i' \in k2} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(Z_{k2} = 1 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \tilde{P}_{i,k} \]
\[ = \eta(1 - \eta) \bar{P}_{ik} \times \]
\[ \sum_{\tilde{y}_{i,k}} \left( \prod_{i' \in k2} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(Z_{k2} = 1 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \right) \]
\[ = \eta(1 - \eta) \bar{P}_{ik} \times \left[ \eta \prod_{i' \in k2} \left( \frac{1}{\tilde{y}_{i,k}} \sum_{\tilde{y}_{i,k} = 0} P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = \tilde{y}_{i,k}) P(\tilde{Y}_{i,k} = \tilde{y}_{i,k}) \right) + \]
\[ \varphi \prod_{i' \in k2} \left( P(Y_{i,k} = y_{i,k} \mid \tilde{Y}_{i,k} = 0)(1 - \tilde{P}_{i,k}) \right) \right], \]

where \( \varphi = 1 - \eta - \delta. \)

5) \( Z_k = 1, Z_{k1} = 1, Z_{k2} = 1: \)

We need to find
\[ \omega_{ik} = P(\tilde{Y}_{ik} = 1 \mid Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_k = 1, Y_{k2} = y_{k2}) \]
\[ = \frac{P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2}, \tilde{Y}_{ik} = 1)}{P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2})}. \]

The denominator can be written as:
\[
P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2}) \\
= \sum_{Y_{k1}, Y_{k2}} P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2} \mid \tilde{Y}_{k1} = \tilde{y}_{k1}) P(\tilde{Y}_{k1} = \tilde{y}_{k1}) \\
= \sum_{Y_{k1}, Y_{k2}} \left\{ P(Z_k = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2}) P(Z_{k1} = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}) \right\} \\
\times P(\tilde{Y}_{k1} = \tilde{y}_{k1}) P(\tilde{Y}_{k2} = \tilde{y}_{k2}) \\
= \sum_{Y_{k1}, Y_{k2}} \left\{ P(Z_k = 1 \mid \tilde{Y}_{k1} = \tilde{y}_{k1}, \tilde{Y}_{k2} = \tilde{y}_{k2}) \prod_{i \in k1} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \\
\times P(\tilde{Y}_{k1} = \tilde{y}_{k1}) P(\tilde{Y}_{k2} = \tilde{y}_{k2}) \\
= (1 - \delta)^3 \prod_{i \in 1} \left( P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 0) (1 - \tilde{p}_{ik}) \right) + \\
\eta^2 (1 - \delta) \prod_{i \in k1} \left( P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 0) (1 - \tilde{p}_{ik}) \right) \\
\times \sum_{y_{ik} = 0} \left\{ \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} + \\
\eta^2 (1 - \delta) \sum_{y_{ik} = 0} \left\{ \prod_{i \in k1} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \\
\times \prod_{i \in k2} \left\{ P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 0) (1 - \tilde{p}_{ik}) \right\} + \\
\eta^3 \sum_{y_{ik} = 0} \left\{ \prod_{i \in k1} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \\
\times \sum_{y_{ik} = 0} \left\{ \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) P(\tilde{Y}_{ik} = \tilde{y}_{ik}) \right\}
\]

and the numerator can be expressed as:

\[
P(\tilde{Y}_{ik} = 1, Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2}) \\
= \sum_{Y_{ik}, Y_{ik}} P(Z_k = 1, Z_{k1} = 1, Y_{k1} = y_{k1}, Z_{k2} = 1, Y_{k2} = y_{k2} \mid \tilde{Y}_{ik} = 1, \tilde{Y}_{-ik} = \tilde{y}_{-ik}) \\
\times P(\tilde{Y}_{ik} = 1, \tilde{Y}_{-ik} = \tilde{y}_{-ik}) \\
= \sum_{Y_{ik}, Y_{ik}} \eta^2 P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \prod_{i \in k1, j \neq i} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \\
\times \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) P(Z_{k2} = 1 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{-ik} = \tilde{y}_{-ik}) \tilde{p}_{ik} \\
= \eta^2 \tilde{p}_{ik} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \times \\
\sum_{y_{ik}, y_{ik}} \left\{ \prod_{i \in k1, j \neq i} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \\
\times P(Z_{k2} = 1 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{-ik} = \tilde{y}_{-ik}) \\
= \eta^2 \tilde{p}_{ik} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \times \sum_{y_{ik}, y_{ik}} \left\{ \prod_{i \in k1, j \neq i} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} \times \\
\sum_{y_{ik}, y_{ik}} \left\{ \prod_{i \in k2} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = \tilde{y}_{ik}) \right\} P(Z_{k2} = 1 \mid \tilde{Y}_{k2} = \tilde{y}_{k2}) P(\tilde{Y}_{-ik} = \tilde{y}_{-ik}) \tilde{p}_{ik}
\[
\eta^2 \tilde{p}_{ik} P(Y_{ik} = y_{ik} \mid \tilde{Y}_{ik} = 1) \sum_{y_{ik} = 1} \left( \prod_{i' \in k, i' \neq i} P(Y_{i'k} = y_{i'k} \mid \tilde{Y}_{i'k} = \tilde{y}_{i'k}) P(\tilde{Y}_{i'k} = \tilde{y}_{i'k}) \right) \times \\
\left[ \eta \prod_{i' \in k} \left( \frac{1}{\tilde{y}_{ik}} \sum_{y_{ik} = 0} P(Y_{i'k} = y_{i'k} \mid \tilde{Y}_{i'k} = \tilde{y}_{i'k}) P(\tilde{Y}_{i'k} = \tilde{y}_{i'k}) \right) + \varphi \prod_{i' \in k} \left( P(Y_{i'k} = y_{i'k} \mid \tilde{Y}_{i'k} = 0)(1 - \tilde{p}_{i'k}) \right) \right].
\]
Appendix C

This appendix includes the derivations of the expression for $\gamma_{ij}$ for the array testing protocol. We can write $\gamma_{ij}$ as

$$\gamma_{ij} = \frac{1}{\prod_{s,t \in Q} P(Y_{st} = y_{st} | \tilde{Y}_{st} = \tilde{y}_{st}) \prod_{i' = 1}^{I} \prod_{j' = 1}^{J} \tilde{p}_{i'j'} (1 - \tilde{p}_{i'j'})^{1-\tilde{y}_{i'j'}}},$$

where we use the same conditional assumptions as in Appendix A and

$$\prod_{s,t \in Q} P(Y_{st} = y_{st} | \tilde{Y}_{st} = \tilde{y}_{st}) \prod_{i' = 1}^{I} \prod_{j' = 1}^{J} \tilde{p}_{i'j'} (1 - \tilde{p}_{i'j'})^{1-\tilde{y}_{i'j'}}$$

denotes the product is taken over all combinations of $i' = 1, \ldots, I$ and $j' = 1, \ldots, J$ except the $(i, j)$ combination. Then

$$\gamma_{ij} = \frac{P(\tilde{Y}_{ij} = \tilde{y}_{ij}, \tilde{Y}_{i'j'} = \tilde{y}_{i'j'}, R = r, C = c, Y_{Q} = y_{Q})}{P(\tilde{Y}_{ij} = \tilde{y}_{ij}, \tilde{Y}_{i'j'} = \tilde{y}_{i'j'}, R = r, C = c, Y_{Q} = y_{Q})} = \frac{P(\tilde{Y}_{ij} = \tilde{y}_{ij}, Y_{Q} = y_{Q}) P(\tilde{Y}_{i'j'} = \tilde{y}_{i'j'}, C = c, Y_{Q} = y_{Q})}{P(\tilde{Y}_{ij} = \tilde{y}_{ij}, Y_{Q} = y_{Q}) P(\tilde{Y}_{i'j'} = \tilde{y}_{i'j'}, C = c)}.$$
\[
\tau = \left[ \prod_{j=1 \atop j \neq i}^{J} P(C_{j'} = c_{j'} \mid \tilde{Y}_{i,j} = \tilde{y}_{i,j}, \ldots, \tilde{Y}_{j,j} = \tilde{y}_{j,j}) \right] \times \left[ \prod_{j=1 \atop j \neq i}^{J} P(C_{j'} = c_{j'} \mid \tilde{Y}_{i,j} = \tilde{y}_{i,j}, \ldots, \tilde{Y}_{j,j} = \tilde{y}_{j,j}) \right] \left[ \prod_{i=1 \atop i \neq j}^{I} \prod_{j=1 \atop j \neq i}^{J} \tilde{p}_{i,j'} (1 - \tilde{p}_{i,j'})^{1-\tilde{y}_{i,j}} \right]
\]

for notational simplicity. Noting that \( \tilde{Y}_{ij} = 1 \) is in the numerator of Equation (28) and if \((i, j) \in Q\), we find that Equation (29) becomes

\[
P(\tilde{Y}_{ij} = 1, \tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q})
= \tau \left[ \prod_{(s,t) \in Q \setminus \{(i,j)\}} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st}) \right] P(R_{i} = r_{i} \mid \tilde{R}_{i} = 1) P(C_{j} = c_{j} \mid \tilde{C}_{j} = 1) \times
P(Y_{ij} = y_{ij} \mid \tilde{Y}_{ij} = 1) \tilde{p}_{ij},
\]

where \((s,t) \in Q \setminus \{(i,j)\}\) means all indices in \(Q\) except for \((i, j)\) and \(\tilde{R}_{i}\) and \(\tilde{C}_{j}\) are the true values for \(R_{i}\) and \(C_{j}\), respectively. When \((i, j) \notin Q\), Equation (29) becomes

\[
P(\tilde{Y}_{ij} = 1, \tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q})
= \tau \left[ \prod_{(s,t) \in Q} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st}) \right] P(R_{i} = r_{i} \mid \tilde{R}_{i} = 1) P(C_{j} = c_{j} \mid \tilde{C}_{j} = 1) \tilde{p}_{ij},
\]

The above equation helps to show the contributions that the individual retests have on the probabilities. Simply, for large sensitivities and specificities, they contribute values close to 0 or 1.

Second, to find the denominator of Equation (28), note that

\[
P(\tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q})
= P(\tilde{Y}_{j} = 0, \tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q}) +
P(\tilde{Y}_{j} = 1, \tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q}).
\]

Using results from Equation (29), we can write the probability for \((i, j) \in Q\) as

\[
P(\tilde{\mathbf{Y}}_{-i,j} = \tilde{\mathbf{y}}_{-i,j}, \mathbf{R} = r, \mathbf{C} = c, \mathbf{Y}_{Q} = y_{Q})
= \tau \left[ \prod_{(s,t) \in Q \setminus \{(i,j)\}} P(Y_{st} = y_{st} \mid \tilde{Y}_{st} = \tilde{y}_{st}) \right] \times
\left[ P(R_{i} = r_{i} \mid \tilde{Y}_{i,j} = \tilde{y}_{i,j}, \ldots, \tilde{Y}_{j,j} = \tilde{y}_{j,j}) \right] \times
\left[ P(R_{i} = r_{i} \mid \tilde{Y}_{i,j} = \tilde{y}_{i,j}, \ldots, \tilde{Y}_{j,j} = \tilde{y}_{j,j}) \right] \times
\]
\[ P(C_j = c_j | \bar{Y}_{ij} = \bar{y}_{ij}, \ldots, \bar{Y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij})P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 0) (1 - \bar{p}_y) + \]
\[ P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 1)\bar{p}_y \]

and for \((i, j) \notin Q:\)
\[ P(\bar{Y}_{-i,-j} = \bar{y}_{-i,-j}, R = r, C = c, Y_Q = y_Q) = \tau \prod_{(s,t) \in Q} P(Y_{st} = y_{st} | \bar{Y}_{st} = \bar{y}_{st}) \{P(R_i = r_i | \bar{Y}_{i1} = \bar{y}_{i1}, \ldots, \bar{Y}_{ij} = \bar{y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij}) (1 - \bar{p}_y) + \]
\[ P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 1)\bar{p}_y \} . \]

Combining all the results, we have for \((i, j) \in Q:\)
\[ \gamma_y = \frac{P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 1)\bar{p}_y}{\{P(R_i = r_i | \bar{Y}_{i1} = \bar{y}_{i1}, \ldots, \bar{Y}_{ij} = \bar{y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij}) \times \]
\[ P(C_j = c_j | \bar{Y}_{ij} = \bar{y}_{ij}, \ldots, \bar{Y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij})P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 0) (1 - \bar{p}_y) + \]
\[ P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)P(Y_{ij} = y_{ij} | \bar{Y}_{ij} = 1)\bar{p}_y \} . \]

and for \((i, j) \notin Q:\)
\[ \gamma_y = \frac{P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)\bar{p}_y}{\{P(R_i = r_i | \bar{Y}_{i1} = \bar{y}_{i1}, \ldots, \bar{Y}_{ij} = \bar{y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij}) \times \]
\[ P(C_j = c_j | \bar{Y}_{ij} = \bar{y}_{ij}, \ldots, \bar{Y}_{ij} = 0, \ldots, \bar{Y}_{ij} = \bar{y}_{ij}) (1 - \bar{p}_y) + \]
\[ P(R_i = r_i | \bar{R}_i = 1)P(C_j = c_j | \bar{C}_j = 1)\bar{p}_y \} . \]

Note that for the case of no individual retests, the formula for \((i, j) \notin Q\) should be used for all \(i\) and \(j\).
Appendix D

This appendix is a web appendix for the Chapter 2 paper submission. We show here a histogram of the true individual probabilities for one simulated data set in Section 2.3.

Figure D.1. Histogram of the true individual probabilities for one simulated data set in Section 2.3.
Appendix E

This appendix is a web appendix for the Chapter 2 paper submission. Below are the average number of tests performed by each protocol for 500 simulated data sets, each containing 5000 individuals, with \( \eta = \delta = 0.99 \) in Section 2.3.2.

Table E.1. Average number of tests performed by each protocol for 500 simulated data sets in Section 2.3.2.

<table>
<thead>
<tr>
<th>Group Size</th>
<th>IG</th>
<th>Dorfman</th>
<th>Halving</th>
<th>Array w/o retesting</th>
<th>Array w/ retesting</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1250</td>
<td>1522</td>
<td>1500</td>
<td>2502</td>
<td>2652</td>
</tr>
<tr>
<td>6</td>
<td>834</td>
<td>1214</td>
<td>1129</td>
<td>1669</td>
<td>1812</td>
</tr>
<tr>
<td>8</td>
<td>625</td>
<td>1111</td>
<td>968</td>
<td>1254</td>
<td>1398</td>
</tr>
<tr>
<td>10</td>
<td>500</td>
<td>1087</td>
<td>891</td>
<td>1000</td>
<td>1144</td>
</tr>
<tr>
<td>12</td>
<td>417</td>
<td>1107</td>
<td>857</td>
<td>837</td>
<td>993</td>
</tr>
<tr>
<td>14</td>
<td>358</td>
<td>1149</td>
<td>848</td>
<td>720</td>
<td>897</td>
</tr>
<tr>
<td>16</td>
<td>313</td>
<td>1196</td>
<td>844</td>
<td>633</td>
<td>831</td>
</tr>
<tr>
<td>18</td>
<td>278</td>
<td>1251</td>
<td>854</td>
<td>564</td>
<td>790</td>
</tr>
<tr>
<td>20</td>
<td>250</td>
<td>1321</td>
<td>875</td>
<td>510</td>
<td>771</td>
</tr>
</tbody>
</table>
Appendix F

This appendix is a web appendix for the Chapter 2 paper submission. We discuss here additional simulations used to reinforce the findings in Section 2.3. We simulate data for each testing protocol according to the model 
\[
\logit(\tilde{p}_{ik}) = \beta_0 + \beta_i x_{ik} \quad (\logit(\tilde{p}_{ij}) = \beta_0 + \beta_i x_{ij} \text{ for the array testing protocols}),
\]
where \( \beta_0 = -6 \) and \( \beta_1 = 4.7 \). The covariates are generated from a Uniform(0, 1) distribution. These configurations provide an overall mean prevalence of about 0.05. The sensitivity and specificity are set to be \( \eta = \delta = 0.99 \). Each simulated data set contains 5000 individuals. The range of the group sizes included in this study is reasonable given the prevalence level.

Figures F.1-F.3 give the results. Overall, we see that the results from Chapter 2 continue to hold true here. Note that \( \psi \) of IG begins to increase with the group sizes in Figure F.3, which it did not for the simulations in Chapter 2. This occurs due to the larger overall prevalence that leads to some of the larger group sizes not being ideal for IG.

Note that Figure F.4 provides a histogram of the true individual probabilities for one simulated data set under the simulation settings.
Figure F.1. Estimated relative efficiencies calculated by Equation (6) based on 500 simulated data sets. Dorfman and halving are compared to IG. Array testing is compared with and without retests.
Figure F.2. Averaged $\text{Var} (\hat{\beta}_1)$ for 500 simulated data sets. The horizontal dashed line corresponds to $\text{Var} (\hat{\beta}_1)$ from individual testing. The right-side plot is the same as on the left-side except we omit IG in order to reduce the y-axis scale.
Figure F.3. Average number of tests per unit of information calculated by Equation (7) based on 500 simulated data sets. Note that $\psi = 903$ for individual testing.

Figure F.4. A histogram of the true individual probabilities for one simulated data set.
Appendix G

This appendix is a web appendix for the Chapter 3 paper submission. In this appendix, we give the proof of Theorem 1 in Section 3.3.2. The covariance between $Z_{jk}$ and $Z_{j'k}$ is

$$
Cov(Z_{jk}, Z_{j'k}) = P(Z_{jk} = 1, Z_{j'k} = 1) - P(Z_{jk} = 1)P(Z_{j'k} = 1),
$$

where

$$
P(Z_{jk} = 1, Z_{j'k} = 1) = \sum_{y_{jk}} \sum_{y_{j'k}} P(Z_{jk} = 1, Z_{j'k} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})P(\tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})
$$

= $\sum_{y_{j'k}} \sum_{y_{jk}} P(Z_{jk} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk})P(Z_{j'k} = 1 | \tilde{Y}_{j'k} = \tilde{y}_{j'k})P(\tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})$ (30)

and $\tilde{Y}_{jk} = (\tilde{Y}_{jk}, ..., \tilde{Y}_{ijk})'$. In Equation (30), we write the joint probability $P(Z_{jk} = 1, Z_{j'k} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})$ as the product of the marginal conditional probabilities $P(Z_{jk} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk})$ and $P(Z_{j'k} = 1 | \tilde{Y}_{j'k} = \tilde{y}_{j'k})$ using a conditional independence assumption (see Litvak, Tu, and Pagano (1994) for justification).

Categorizing with respect to the possible values of $\tilde{y}_{jk}$ and $\tilde{y}_{j'k}$, we split the summation in Equation (30) into four parts. The first part corresponds to the case where $\tilde{y}_{jk}$ and $\tilde{y}_{j'k}$ are vectors of 0’s, which leads to

$$
P(Z_{jk} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk})P(Z_{j'k} = 1 | \tilde{Y}_{j'k} = \tilde{y}_{j'k})P(\tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})
$$

$$
= (1 - \delta_{j})(1 - \delta_{j'})\prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0).
$$

The second part corresponds to the case where $\tilde{y}_{j'k}$ is a vector of 0’s and at least one $\tilde{y}_{ijk}$ for $1 \leq i \leq I_k$ is not 0. Defining $Q_j = \{\tilde{y}_{jk} | \text{at least one } \tilde{y}_{ijk} \text{ for } 1 \leq i \leq I_k \text{ is not 0}\}$, we obtain
Because $\sum_{y_{jk} \in Q_j} P(\tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = 0) + P(\tilde{Y}_{jk} = 0, \tilde{Y}_{j'k} = 0) = P(\tilde{Y}_{jk} = 0)$, Equation (31) can be further written as

$$\eta_j (1 - \delta_j) \left\{ \prod_{i=1}^{I_k} (1 - \tilde{p}_{ij'k}) - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) \right\}.$$ 

Finally, the fourth part corresponds to the case where there is at least one positive individual for both disease $j$ and $j'$. We find that

$$\sum_{y_{jk} \in Q_j} \sum_{y_{j'k} \in Q_{j'}} P(Z_{jk} = 1, Z_{j'k} = 1 | \tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k}) P(\tilde{Y}_{jk} = \tilde{y}_{jk}, \tilde{Y}_{j'k} = \tilde{y}_{j'k})$$

$$= \eta_j \eta_{j'} \left\{ 1 - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) \right\}$$

$$- \sum_{\tilde{y}_{jk} \in Q_j} P(\tilde{Y}_{jk} = 0, \tilde{Y}_{j'k} = \tilde{y}_{j'k})$$

$$= \eta_j \eta_{j'} \left\{ 1 - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) \right\}$$

$$- \left\{ \prod_{i=1}^{I_k} (1 - \tilde{p}_{ij'k}) - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) \right\}$$

$$- \left\{ \prod_{i=1}^{I_k} (1 - \tilde{p}_{ij'k}) - \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) \right\}$$

$$= \eta_j \eta_{j'} \left\{ 1 + \prod_{i=1}^{I_k} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{i'jk} = 0) - \prod_{i=1}^{I_k} (1 - \tilde{p}_{ij'k}) - \prod_{i=1}^{I_k} (1 - \tilde{p}_{ij'k}) \right\}.$$
\[ P(Z_{jk} = 1, Z_{jk} = 1) = (1 - \delta_j)(1 - \delta_{j'}) \prod_{i=1}^{l_k} P(Y_{ijk} = 0, Y_{ijk} = 0) + \eta_j(1 - \delta_{j'}) \left\{ \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) - \prod_{i=1}^{l_k} P(Y_{ijk} = 0, Y_{ijk} = 0) \right\} \\
+ \eta_{j'}(1 - \delta_j) \left\{ \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) - \prod_{i=1}^{l_k} P(Y_{ijk} = 0, Y_{ijk} = 0) \right\} \\
+ \eta_j\eta_{j'} \left\{ 1 + \prod_{i=1}^{l_k} P(Y_{ijk} = 0, Y_{ijk} = 0) - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) \right\} . \]

Shifting focus to \( P(Z_{jk} = 1)P(Z_{jk} = 1) \) in Equation (30), we notice that

\[ P(Z_{jk} = 1) = \sum_{y_{jk}} P(Z_{jk} = 1 \mid Y_{jk} = \tilde{y}_{jk}) P(\tilde{Y}_{jk} = \tilde{y}_{jk}) \\
= (1 - \delta_j) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) + \eta_j \left\{ 1 - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) \right\} . \]

Thus,

\[ P(Z_{jk} = 1)P(Z_{jk} = 1) = \left\{ (1 - \delta_j) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) + \eta_j \left\{ 1 - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) \right\} \right\} \times \]

\[ \left\{ (1 - \delta_{j'}) \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) + \eta_{j'} \left\{ 1 - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) \right\} \right\} \\
= (1 - \delta_j)(1 - \delta_{j'}) \left\{ \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ijk}) \right\} \\
+ \eta_j(1 - \delta_{j'}) \left\{ \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ijk}) \right\} \\
+ \eta_{j'}(1 - \delta_j) \left\{ \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ijk}) \right\} \\
+ \eta_j\eta_{j'} \left\{ 1 - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) - \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk}) + \prod_{i=1}^{l_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ijk}) \right\} . \]

Subtracting \( P(Z_{jk} = 1)P(Z_{jk} = 1) \) from \( P(Z_{jk} = 1, Z_{jk} = 1) \), the covariance becomes
\begin{align*}
\text{Cov}(Z_{jk}, Z_{j'k}) &= (1 - \delta_j)(1 - \delta_{j'}) \left\{ \prod_{i=1}^{h_j} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0) - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} \\
&\quad - \eta_j(1 - \delta_j) \left\{ \prod_{i=1}^{h_j} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0) - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} \\
&\quad - \eta_{j'}(1 - \delta_{j'}) \left\{ \prod_{i=1}^{h_{j'}} P(\tilde{Y}_{ij'k} = 0) - \prod_{i=1}^{h_{j'}} (1 - \tilde{p}_{ij'k})(1 - \tilde{p}_{ij'k}) \right\} \\
&\quad + \eta_j \eta_{j'} \left\{ \prod_{i=1}^{h_j} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0) - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} \\
&= (\delta_j + \eta_j - 1)(\delta_{j'} + \eta_{j'} - 1) \left\{ \prod_{i=1}^{h_j} P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0) - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\}.
\end{align*}

Because \( P(\tilde{Y}_{ijk} = 0, \tilde{Y}_{ij'k} = 0) = \text{Cov}(1 - \tilde{Y}_{ijk}, 1 - \tilde{Y}_{ij'k}) \), we can write

\begin{align*}
\text{Cov}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) &= (\delta_j + \eta_j - 1)(\delta_{j'} + \eta_{j'} - 1) \times \\
&\quad \left\{ \prod_{i=1}^{h_j} \text{Cov}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) + (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \\
&= (\delta_j + \eta_j - 1)(\delta_{j'} + \eta_{j'} - 1) \\
&\quad \left\{ \prod_{i=1}^{h_j} \left[ \sqrt{\text{Var}(\tilde{Y}_{ijk})\text{Var}(\tilde{Y}_{ij'k})} + \text{Corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) \right] + (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} - \prod_{i=1}^{h_j} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k})
\end{align*}

for \( 1 \leq j, j' \leq J, j \neq j' \), and \( k = 1, \ldots, K \).
Appendix H

This appendix is a web appendix for the Chapter 3 paper submission. The purpose here is to illustrate the existence of a unique solution $\hat{\alpha}$ and the consistency of $\hat{\alpha}$ when $\beta$ is known, as discussed in Section 3.3.2. For generality, we examine an unspecified working correlation structure. More refined arguments for specific correlation structures follow analogously.

For the same group, but different disease, Theorem 1 gives the covariance between two group responses as

$$
\text{Cov}(Z_{jk}, Z_{j'k}) = (\delta_j + \eta_j - 1)(\delta_{j'} + \eta_{j'} - 1) \times
\left[ \prod_{i=1}^l \left\{ \alpha_{j'} \sqrt{\frac{\tilde{p}_{j'k}}{1 - \tilde{p}_{j'k}}} \tilde{p}_{j'k} (1 - \tilde{p}_{j'k}) + (1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k}) \right\} \right. \\
\left. - \prod_{i=1}^l (1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k}) \right],
$$

(32)

which is a function of $\text{Corr}(\tilde{Y}_{jk}, \tilde{Y}_{j'k}) = \alpha_{j'}$, for $j \neq j'$. It is obvious that this function passes through the origin and has positive coefficients. Therefore, $\text{Cov}(Z_{jk}, Z_{j'k})$ is an increasing function of $\alpha_{j'}$ when $\alpha_{j'} \geq 0$.

For $-1 \leq \alpha_{j'} < 0$, notice that $\alpha_{j'} \sqrt{\frac{\tilde{p}_{j'k}}{1 - \tilde{p}_{j'k}}} \tilde{p}_{j'k} (1 - \tilde{p}_{j'k}) + (1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k})$ is an increasing function of $\alpha_{j'}$, for each $i = 1, \ldots, I_k$. When $\tilde{p}_{j'k}$ and $\tilde{p}_{j'k}$ are less than or equal to 0.5, we have

$$
(1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k}) + \alpha_{j'} \sqrt{\frac{\tilde{p}_{j'k}}{1 - \tilde{p}_{j'k}}} \tilde{p}_{j'k} (1 - \tilde{p}_{j'k}) \\
\geq (1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k}) - \sqrt{\frac{\tilde{p}_{j'k}}{1 - \tilde{p}_{j'k}}} \tilde{p}_{j'k} (1 - \tilde{p}_{j'k}) \\
= \sqrt{(1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k})} \left\{ \sqrt{(1 - \tilde{p}_{j'k})(1 - \tilde{p}_{j'k})} - \sqrt{\tilde{p}_{j'k} \tilde{p}_{j'k}} \right\} \\
\geq 0
$$
Hence, \( \alpha_{ij'} \sqrt{\tilde{p}_{ijk}(1-\tilde{p}_{ijk})\tilde{p}_{j'k}(1-\tilde{p}_{j'k})} + (1-\tilde{p}_{ijk})(1-\tilde{p}_{j'k}) \) is non-negative in this situation, and the product

\[
\prod_{i=1}^{l} \left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk}(1-\tilde{p}_{ijk})\tilde{p}_{j'k}(1-\tilde{p}_{j'k})} + (1-\tilde{p}_{ijk})(1-\tilde{p}_{j'k}) \right\}
\]

is an increasing function of \( \alpha_{ij'} \), for \(-1 \leq \alpha_{ij'} < 0 \). Thus, \( \text{Cov}(Z_{jk}, Z_{j'k}) \) is an increasing function of \( \alpha_{ij'} \) for \( \alpha_{ij'} \geq 0 \) and also for \(-1 \leq \alpha_{ij'} < 0 \) when \( \tilde{p}_{ijk} \) and \( \tilde{p}_{j'k} \) are less than or equal to 0.5. The right hand side of Equation (14) is the sum of \( \text{Cov}(Z_{jk}, Z_{j'k}) \) terms over \( k = 1, \ldots, K \), and thus will also be monotonic increasing in \( \alpha_{ij'} \). The existence of a unique solution in Equation (14) immediately follows.

When some \( \tilde{p}_{ijk} \) and \( \tilde{p}_{j'k} \) are greater than 0.5 and \(-1 \leq \alpha_{ij'} < 0 \), \( \text{Cov}(Z_{jk}, Z_{j'k}) \) may not be monotone, which could lead to more than one solution in Equation (14). However, this would be unlikely to occur for the following reasons: (a) Disease statuses most likely will have a non-negative correlation, i.e., \( \alpha_{ij'} \geq 0 \), (b) Most individual probabilities will be less than 0.5 in realistic settings; otherwise, we would not use group testing, and (c) The estimate of \( \alpha_{ij'} \) comes about through solving Equation (14), where we sum over all groups rather than examine only one group. Because most \( \tilde{p}_{ijk} \) and \( \tilde{p}_{j'k} \) are less than 0.5 within groups, we postulate that the right hand side of (14) will still be a monotonic function, which leads to a unique solution for \( \alpha_{ij'} \).

To prove the consistency of \( \hat{\alpha} \), we define

\[
g_k(\alpha_{ij'}) = (Z_{jk} - \theta_{jk})(Z_{j'k} - \theta_{j'k}) - (\delta_k + \eta_k - 1)(\delta_{j'} + \eta_{j'} - 1) \times
\[
\prod_{i=1}^{l} \left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk}(1-\tilde{p}_{ijk})\tilde{p}_{j'k}(1-\tilde{p}_{j'k})} + (1-\tilde{p}_{ijk})(1-\tilde{p}_{j'k}) \right\}
\]

\[
-\prod_{i=1}^{l} (1-\tilde{p}_{ijk})(1-\tilde{p}_{j'k})
\]
for \( k = 1, \ldots, K, \ j \neq j' \), and \( g_K = \frac{\sum_{k=1}^{K} g_k}{K} \) for a fully unspecified working correlation structure. Due to the independence of group responses across groups, each \( g_k \) is independent of each other. Furthermore, clearly \( E(g_k) = 0 \) from Equation (32), so that \( E(g_k) = 0 \) as well. It is obvious that \( \text{Var}(g_k) < \infty \) and

\[
\lim_{k \to \infty} \frac{\text{Var}(\sum_{k=1}^{K} g_k)}{K^2} = \lim_{k \to \infty} \frac{\sum_{k=1}^{K} \text{Var}(g_k)}{K^2} < \infty.
\]

By the Chebyshev law of large numbers (e.g., see Serfling (1980, p. 27)), \( g_K(\alpha_{j'}) \to 0 \) in probability. Similar to Liang and Zeger (1995, p. 163), \( \hat{\alpha}_{j'} = g_K^{-1}(0) \to \alpha_{j'} \) provided that \( g_K \) is continuous and one-to-one. Thus \( \hat{\alpha} \) is a consistent estimator for \( \alpha \) when \( \beta \) is known.
Appendix I

This appendix is a web appendix for the Chapter 3 paper submission. The purpose here is to show how to derive the linear and quadratic coefficients in Section 3.3.2 and explain why using the first and second order terms give close estimates to the exact solution in Equation (14). To find the linear term in

$$\prod_{i=1}^{I} \left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk} (1 - \tilde{p}_{ijk}) \tilde{p}_{j'k} (1 - \tilde{p}_{j'k})} + (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \right\},$$

we choose a $$\alpha_{ij'} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk}) (1 - \tilde{p}_{j'k})}$$ term and multiply it with the other $$I_k - 1$$ remaining $$(1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k})$$ terms to obtain

$$\begin{align*}
\left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk}) (1 - \tilde{p}_{j'k})} \right\} \prod_{i=1}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) + \\
\left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk}) (1 - \tilde{p}_{j'k})} \right\} \prod_{i=2}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) + \\
... + \\
\left\{ \alpha_{ij'} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk}) (1 - \tilde{p}_{j'k})} \right\} \prod_{i=t_k}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \\
= \alpha_{ij'} \sum_{i=1}^{I_k} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk}) (1 - \tilde{p}_{j'k})} \prod_{i'=1}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \\
= \alpha_{ij'} \left\{ \prod_{i=1}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \right\} \sum_{i=1}^{I_k} \sqrt{\frac{\tilde{p}_{ijk} \tilde{p}_{j'k}}{(1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k})}}.
\end{align*}$$

Similarly for the quadratic term, from the $$I_k$$ products we pick two $$\alpha_{ij'} \sqrt{\tilde{p}_{ijk} \tilde{p}_{j'k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k})}$$ terms and multiply them with the other $$I_k - 2$$ remaining $$(1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k})$$ terms. The product can be written as

$$\alpha_{ij'}^2 \left\{ \prod_{i=1}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \right\} \sum_{1 \leq k < \ell \leq I_k} \sqrt{\frac{\tilde{p}_{h_{jk}h_{j'k}}}{(1 - \tilde{p}_{h_{jk}})(1 - \tilde{p}_{h_{j'k}})}} \sqrt{\frac{\tilde{p}_{h_{jk}h_{j'k}}}{(1 - \tilde{p}_{h_{jk}})(1 - \tilde{p}_{h_{j'k}})}}.$$

We can obtain the coefficient of the $$l^{th}$$ order term as

$$\alpha_{ij'}^l \left\{ \prod_{i=1}^{I} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{j'k}) \right\} \sum_{1 \leq k < \ldots < \ell \leq I_k} \prod_{m=1}^{l} \sqrt{\frac{\tilde{p}_{h_{jk}h_{j'k}}}{(1 - \tilde{p}_{h_{jk}})(1 - \tilde{p}_{h_{j'k}})}}.$$
for \( l = 1, \ldots, I_k \). Because the \( \hat{p}_{ijk} \) are generally small for group testing applications, we expect that \( \prod_{m=1}^{l} \sqrt{\hat{p}_{m,jk} \hat{p}_{m,j^*k}} / \{(1 - \hat{p}_{m,jk})(1 - \hat{p}_{m,j^*k})\} \) becomes extremely small rather quickly as \( l \) increases. For example, in the simplified case of \( \hat{p}_{ijk} \equiv \hat{p} \), one can show that the ratio of the \( l \text{th} \) to the \((l - 1)\text{th}\) term is

\[
\left\{ \begin{array}{c}
I_k \\
l
\end{array} \right\} \frac{\hat{p}}{1 - \hat{p}} \alpha_{jj'} = \left\{ \begin{array}{c}
I_k \\
l
\end{array} \right\} \frac{I_k - l + 1}{l} \frac{\hat{p}}{1 - \hat{p}} \alpha_{jj'}.
\]  (33)

Because \((I_k - l + 1) / l\) is a decreasing function of \( l \), \( \hat{p} \) is generally small, and \(-1 < \alpha_{jj'} < 1\), Equation (33) is close to 0 and decreases as \( l \) increases. For example, if \( \hat{p} = 0.03 \), \( I_k = 5 \), and \( \alpha_{jj'} = 0.6 \), then the quadratic term is 3.7\% of the linear term and the cubic term is only 1.9\% of the quadratic term.

To examine the approximation more closely, we performed the following simulation study. Consider the model

\[
\logit(\hat{p}_{ijk}) = \beta_{0j} + \beta_{1j} x_{1ik},
\]

where \( j = 1, 2 \) and \( Corr(\tilde{Y}_{1ik}, \tilde{Y}_{12k}) = \alpha \). The covariate \( x_{1ik} \) is generated from a \( \text{gamma}(17, 1.4) \) distribution. The true parameters of the model are \( \beta_{01} = -7 \), \( \beta_{11} = 0.13 \), \( \beta_{02} = -6 \), and \( \beta_{12} = 0.1 \). We simulate 5 data sets and estimate \( \alpha \) using Equation (14) (we refer to this as “exact”) and using first- and second-order approximations as outlined in Section 3.3.2. Table I.1 gives the estimates, and it shows that the second-order approximation works especially well at approximating the exact result.
Table I.1. Comparison of $\alpha$ estimates averaged over 5 simulated data sets. The average time in minutes that our R function took to estimate each model is shown for a computer with a 2.2GHZ processor and 3GB of memory.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\eta_j$</th>
<th>$\delta_j$</th>
<th>$K$</th>
<th>$I$</th>
<th>Exact $\hat{\alpha}$</th>
<th>Exact Time</th>
<th>First-order $\hat{\alpha}$</th>
<th>First-order Time</th>
<th>Second-order $\hat{\alpha}$</th>
<th>Second-order Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>1000</td>
<td>5</td>
<td>0.6262</td>
<td>1.25</td>
<td>0.6486</td>
<td>0.47</td>
<td>0.6264</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>10</td>
<td>0.6072</td>
<td>4.53</td>
<td>0.6551</td>
<td>0.47</td>
<td>0.6087</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>1000</td>
<td>5</td>
<td>0.5700</td>
<td>1.68</td>
<td>0.5899</td>
<td>0.65</td>
<td>0.5701</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>10</td>
<td>0.5684</td>
<td>4.76</td>
<td>0.6026</td>
<td>0.52</td>
<td>0.5695</td>
<td>0.85</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1</td>
<td>1000</td>
<td>5</td>
<td>0.1715</td>
<td>1.20</td>
<td>0.1733</td>
<td>0.43</td>
<td>0.1715</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>10</td>
<td>0.1767</td>
<td>4.58</td>
<td>0.1807</td>
<td>0.51</td>
<td>0.1768</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.95</td>
<td>1000</td>
<td>5</td>
<td>0.2190</td>
<td>1.38</td>
<td>0.2217</td>
<td>0.57</td>
<td>0.2191</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>10</td>
<td>0.1835</td>
<td>4.88</td>
<td>0.1883</td>
<td>0.60</td>
<td>0.1836</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Appendix J

This appendix is a web appendix for the Chapter 3 paper submission.

1. Histograms of the true individual probabilities for one simulated data set

Figure J.1. Histograms of the true individual probabilities for one Section 3.4 simulated data set using the model in Equation (16) with $\alpha = 0.6$, $K = 1000$, and $I_k = 5$. 
2. Simulation results from estimating $\beta_{21}$ and $\beta_{22}$ separately

Table J.1. Relative efficiency of the variance estimates for the model in Equation (16). Note that each $\beta_{21}$ and $\beta_{22}$ is estimated separately by the ES algorithm.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$K$</th>
<th>$I_k$</th>
<th>$\hat{\beta}_{01}$</th>
<th>$\hat{\beta}_{02}$</th>
<th>$\hat{\beta}_{11}$</th>
<th>$\hat{\beta}_{12}$</th>
<th>$\hat{\beta}_{21}$</th>
<th>$\hat{\beta}_{22}$</th>
<th>logit($\hat{p}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>2000</td>
<td>5</td>
<td>1.053</td>
<td>1.039</td>
<td>1.054</td>
<td>1.038</td>
<td>1.041</td>
<td>1.047</td>
<td>1.093</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.064</td>
<td>1.069</td>
<td>1.060</td>
<td>1.068</td>
<td>1.061</td>
<td>1.072</td>
<td>1.110</td>
</tr>
<tr>
<td>0.2</td>
<td>2000</td>
<td>5</td>
<td>1.016</td>
<td>1.025</td>
<td>1.014</td>
<td>1.027</td>
<td>1.014</td>
<td>1.025</td>
<td>1.011</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.043</td>
<td>1.043</td>
<td>1.031</td>
<td>1.056</td>
<td>1.041</td>
<td>1.035</td>
<td>1.024</td>
</tr>
</tbody>
</table>
3. Additional simulations

Additional simulations were performed to support the findings in Section 3.4 with different models. We used the model \( \logit(p_{ijk}) = \beta_{0j} + \beta_{1j}x_{ik} \) with \( \beta_{01} = -6.3, \beta_{02} = -6.6, \beta_{11} = 4.0, \beta_{12} = 4.7 \), where \( x_{ik} \sim \text{uniform}(0, 1) \) and \( \eta_j = \delta_j = 1 \) for \( j = 1, 2 \). These covariate and parameter configurations lead to a mean prevalence of approximately 0.02 for disease \( j = 1 \) and 0.03 for disease \( j = 2 \). The results are given in Table J.2, and they are very similar to Table J.1 for the two-covariate model. Specifically, the ES algorithm leads to more efficient estimators than those from separate models; also, further benefits are realized by estimating the correlation between the disease responses. As would be expected, the benefits from estimating the correlation decrease as a function of \( \alpha \).

Table J.2. Relative efficiency of the variance estimates for \( \logit(p_{ijk}) = \beta_{0j} + \beta_{1j}x_{ik} \).

The relative efficiency involving \( \logit(\hat{p}) \) is calculated at values of \( x_{ik} = 0.1, 0.2, \ldots, 0.9 \), and the maximum and minimum relative efficiencies are reported in the table.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( K )</th>
<th>( I_k )</th>
<th>( \hat{\beta}_{01} )</th>
<th>( \hat{\beta}_{02} )</th>
<th>( \hat{\beta}_{11} )</th>
<th>( \hat{\beta}_{12} )</th>
<th>( \logit(\hat{p}) ) max</th>
<th>( \logit(\hat{p}) ) min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1000</td>
<td>5</td>
<td>1.060</td>
<td>1.052</td>
<td>1.059</td>
<td>1.052</td>
<td>1.059</td>
<td>1.025</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.056</td>
<td>1.058</td>
<td>1.048</td>
<td>1.055</td>
<td>1.111</td>
<td>1.036</td>
</tr>
<tr>
<td>0.5</td>
<td>1000</td>
<td>5</td>
<td>1.048</td>
<td>1.044</td>
<td>1.048</td>
<td>1.044</td>
<td>1.048</td>
<td>1.011</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.033</td>
<td>1.021</td>
<td>1.031</td>
<td>1.021</td>
<td>1.037</td>
<td>1.017</td>
</tr>
<tr>
<td>0.2</td>
<td>1000</td>
<td>5</td>
<td>1.043</td>
<td>1.036</td>
<td>1.043</td>
<td>1.037</td>
<td>1.043</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10</td>
<td>1.027</td>
<td>1.018</td>
<td>1.026</td>
<td>1.018</td>
<td>1.027</td>
<td>1.009</td>
</tr>
</tbody>
</table>
Appendix K

This appendix is a web appendix for the Chapter 3 paper submission.

Table K.1. Parameter estimates and estimated standard errors (in parentheses) for the parsimonious model described in Section 3.5. The GEE column corresponds to a model fit to the individual responses using GEE methodology.

<table>
<thead>
<tr>
<th>Term</th>
<th>Disease</th>
<th>ES</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>-5.76(0.61)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>-0.53(0.42)</td>
<td>-</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>-0.03(0.02)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>-0.11(0.02)</td>
<td>-</td>
</tr>
<tr>
<td>Race level #1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>1.84(0.34)</td>
<td>1.31(0.17)</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>0.58(0.20)</td>
<td>0.40(0.10)</td>
</tr>
<tr>
<td>Race level #2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>0.75(0.98)</td>
<td>0.71(0.34)</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>1.06(0.24)</td>
<td>0.69(0.14)</td>
</tr>
<tr>
<td>Race level #3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>0.45(0.93)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>0.04(0.40)</td>
<td>0.06(0.15)</td>
</tr>
<tr>
<td>Symptoms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>1.24(0.42)</td>
<td>0.95(0.17)</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>0.40(0.17)</td>
<td>0.28(0.08)</td>
</tr>
<tr>
<td>Cervical friability</td>
<td>Both</td>
<td>0.27(0.28)</td>
<td>0.10(0.16)</td>
</tr>
<tr>
<td>Pelvic inflammatory</td>
<td>Both</td>
<td>0.63(0.55)</td>
<td>0.62(0.34)</td>
</tr>
<tr>
<td>Cervicitis</td>
<td>Both</td>
<td>0.50(0.18)</td>
<td>0.58(0.10)</td>
</tr>
<tr>
<td>Multiple partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>1.22(0.30)</td>
<td>1.04(0.17)</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>0.27(0.22)</td>
<td>0.47(0.10)</td>
</tr>
<tr>
<td>New partner</td>
<td>Both</td>
<td>0.11(0.18)</td>
<td>-</td>
</tr>
<tr>
<td>Contact to a STD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gonorrhea</td>
<td>1.33(0.29)</td>
<td>1.17(0.18)</td>
</tr>
<tr>
<td></td>
<td>Chlamydia</td>
<td>0.59(0.21)</td>
<td>0.94(0.10)</td>
</tr>
</tbody>
</table>