BIOMETRICS NA, 1–20 DOI: NA
NA NA

Marginal Regression Models for Multiple-Disease Group Testing Data

Boan Zhang* **and Christopher R. Bilder****

Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, U.S.A.

**email:* boan.zhang@huskers.unl.edu

***email:* chris@chrisbilder.com

and

Joshua M. Tebbs

Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

email: tebbs@stat.sc.edu

SUMMARY: Group testing, where individual specimens are composited into groups 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 and to prenatal infectious disease screening data from Kenya.

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

This paper has been submitted for consideration for publication in *Biometrics*

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 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 (2012), opportunistic chlamydia and gonorrhea testing in medical clinics (Gaydos, 2005), and Bovine Viral Diarrhea virus detection in the cattle industry (Munoz-Zanzi et al., 2006). Group testing has also proven to be beneficial in other areas including 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 overall disease prevalence for a population. More recently, this research has shifted towards incorporating 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 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 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, and Delaigle and Meister (2011) develop nonparametric modeling approaches. Group testing

regression models even have been used to detect model misspecification with individual response data, as shown 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 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 (IPP), the Nebraska Public Health Laboratory (NPHL) screens thousands of individuals per year using the GenProbe Aptima Combo 2 assay which tests 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 this problem in the group testing literature, and they do so by estimating overall population prevalences under the assumption that diagnostic tests are perfect.

In our paper, we develop new group testing regression methods for analyzing multiple-disease screening data with imperfect diagnostic tests. Our research deals with modeling correlated binary data, but with the unique aspect that disease responses for each individual are unobserved. Broadly speaking, our paper can be viewed as a generalization of Vansteelandt et al. (2000) and Xie (2001) to model multiple-disease statuses and as 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 2 defines notation and states the model of interest. Section 3 shows how the expectation-solution (ES) algorithm (Elashoff and Ryan, 2004) can be used to model multiple-disease statuses with group testing responses. In

addition, we develop a novel approach to estimate a working correlation structure among the unobserved individual responses by using the observed group responses. Section 4 presents simulation evidence demonstrating that our proposed estimators are consistent and that large-sample inference procedures confer nominal levels. Section 5 applies this work to two disease screening data sets, one from the NPHL and one from a prenatal infectious disease study in Kenya. Finally, Section 6 summarizes this work and suggests future areas of research.

2. Notation and Model

Let $\tilde{Y}_{ijk} = 1$ (0) if individual i in group k is truly positive (negative) for disease j, for $i = 1, \ldots, I_k$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. We assume that $\tilde{\boldsymbol{Y}}_{ik} = (\tilde{Y}_{i1k}, \ldots, \tilde{Y}_{iJk})'$ are independent random vectors across i and k and that $\tilde{Y}_{i1k}, \ldots, \tilde{Y}_{iJk}$ are possibly correlated across j. Let $Z_{jk} = 1$ (0) if group k tests positive (negative) for disease j. 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 should account for this uncertainty. For disease j, define the group test sensitivity and specificity as $\eta_j = P(Z_{jk} = 1|\tilde{Z}_{jk} = 1)$ and $\delta_j = P(Z_{jk} = 0|\tilde{Z}_{jk} = 0)$, respectively, where $\tilde{Z}_{jk}$ denotes the true group binary status for disease j and group k. We assume η_j and δ_j are known for each disease and are not dependent on pool sizes or covariates; 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 (2002) for individual testing regression models.

With covariates $\boldsymbol{x}_{ik} = (x_{1ik}, \ldots, x_{p-1,ik})'$ measured on each individual, our goal is to estimate $P(\tilde{Y}_{ijk} = 1|\boldsymbol{x}_{ik}) \equiv \tilde{p}_{ijk}$ when only the group responses Z_{jk} are available, similar to Vansteelandt et al. (2000) with single-disease models. In all subsequent expectations written in our paper, we condition on the full set of covariates $\boldsymbol{x}_{ik}$ as we did for $\tilde{p}_{ijk}$, but we suppress

this specification for notational simplicity. We consider models of the form

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

where $f(\cdot)$ is a known monotonic, differentiable function and β_{rj} $(r = 0, \ldots, p-1, j = 1, \ldots, J)$ is a regression parameter. This model allows us to estimate the regression parameters jointly for all diseases rather than separately as would be done with J single-disease models.

A joint (or "multiple-disease") model provides distinct advantages over using J single-disease models. First, a joint model enables one to model group testing data as it naturally arises from multiple-disease screening assays. Second, a joint model can incorporate within-subject correlation across the J diseases, unlike single-disease models which essentially ignore this information. By incorporating this correlation, joint modeling leads to more efficient estimators. Finally, a joint model allows for inference across diseases; this enables one to assess whether specific covariates have similar effects across the J disease statuses. Each of these advantages is illustrated in subsequent sections of this paper.

3. Expectation-Solution Algorithm

We use the ES Algorithm to estimate the parameters in Equation (1). The ES algorithm, introduced by Elashoff and Ryan (2004), is a generalization of the expectation-maximization (EM) algorithm given by Dempster, Laird, and Rubin (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 expected values into complete-data estimating equations and solves the equations for the regression 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 (1) using the observed group responses. As we demonstrate shortly, this application of the ES algorithm requires additional work in order to estimate the correlation among the unobserved individual responses.

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 (e.g., see Hardin and Hilbe (2003)) is used to estimate the model in Equation (1). Let $\boldsymbol{R}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_S)$, denote the $J \times J$ working correlation matrix for the individual responses. Define $Cov(\tilde{\boldsymbol{Y}}_{ik}) = \boldsymbol{V}_{ik} = \boldsymbol{B}_{ik}^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{B}_{ik}^{1/2}$ where $\boldsymbol{B}_{ik} = Diag\left(\tilde{p}_{ijk}(1-\tilde{p}_{ijk})\right)$. The estimating equations are

$$\Psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_k \sum_i \Psi_{ik}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_k \sum_i \boldsymbol{D}_{ik}' \boldsymbol{V}_{ik}^{-1}(\tilde{\boldsymbol{y}}_{ik} - \tilde{\boldsymbol{p}}_{ik}) = \boldsymbol{0}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{p-1,1}, \beta_{02}, \ldots, \beta_{p-1,J})'$, $\boldsymbol{D}_{ik} = (\partial/\partial\boldsymbol{\beta})\tilde{\boldsymbol{p}}_{ik}$, $\tilde{\boldsymbol{p}}_{ik} = (\tilde{p}_{i1k}, \ldots, \tilde{p}_{iJk})'$, $\tilde{\boldsymbol{y}}_{ik}$ is a realization of $\tilde{\boldsymbol{Y}}_{ik}$, $\boldsymbol{0}$ is a $pJ \times 1$ vector of 0's, and $\Psi_{ik}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{D}_{ik}' \boldsymbol{V}_{ik}^{-1}(\tilde{\boldsymbol{y}}_{ik} - \tilde{\boldsymbol{p}}_{ik})$ 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 $\boldsymbol{\alpha}$ and solving Equation (2) for $\boldsymbol{\beta}$ in an iterative manner until convergence.

Because the individual responses $\tilde{Y}_{ijk}$ are not observed, we can not use standard GEE methodology. However, analogous to the use of the EM algorithm described in Xie (2001) for a single disease, we can replace the individual responses in Equation (2) by their expected values, conditional on the group responses $\boldsymbol{Z} = (Z_{11}, \ldots, Z_{JK})'$. Because $\tilde{Y}_{ijk}$ is dependent only on its corresponding group response, it suffices to calculate $E(\tilde{Y}_{ijk}|Z_{jk} = 1) = \eta_j \tilde{p}_{ijk}/\theta_{jk}$ and $E(\tilde{Y}_{ijk}|Z_{jk} = 0) = (1-\eta_j)\tilde{p}_{ijk}/(1-\theta_{jk})$, where $\theta_{jk} \equiv P(Z_{jk} = 1)$ is

$$\begin{aligned}\theta_{jk} &= P(Z_{jk}=1|\tilde{Z}_{jk}=1)P(\tilde{Z}_{jk}=1)+P(Z_{jk}=1|\tilde{Z}_{jk}=0)P(\tilde{Z}_{jk}=0)\\ &= \eta_j+(1-\delta_j-\eta_j)\prod_{i=1}^{I_k}(1-\tilde{p}_{ijk}).\end{aligned} \tag{3}$$

When replacing $\tilde{y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$, Equation (2) becomes

$$\Psi^{obs}(\boldsymbol{\beta},\boldsymbol{\alpha})=\sum_k\sum_i\Psi_{ik}^{obs}(\boldsymbol{\beta},\boldsymbol{\alpha})=\sum_k\sum_i\boldsymbol{D}_{ik}'\boldsymbol{V}_{ik}^{-1}(\boldsymbol{\omega}_{ik}-\tilde{\boldsymbol{p}}_{ik})=\mathbf{0}, \tag{4}$$

where $\boldsymbol{\omega}_{ik}=(E(\tilde{Y}_{i1k}|z_{1k}),\ldots,E(\tilde{Y}_{iJk}|z_{Jk}))'$ and $\Psi_{ik}^{obs}(\boldsymbol{\beta},\boldsymbol{\alpha})=\boldsymbol{D}_{ik}'\boldsymbol{V}_{ik}^{-1}(\boldsymbol{\omega}_{ik}-\tilde{\boldsymbol{p}}_{ik})$.

The ES Algorithm successively estimates $\boldsymbol{\alpha}$ and solves Equation (4) for $\boldsymbol{\beta}$ in an iterative manner to obtain parameter estimates, as is done with standard GEE methodology. The initial estimate of $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$. Estimating $\boldsymbol{\alpha}$ at each iteration is not as straightforward as in a standard GEE situation, so we discuss it separately in the next subsection. The final iterative solution to Equation (4) at convergence is the estimate of $\boldsymbol{\beta}$, which we denote by $\hat{\boldsymbol{\beta}}$.

3.2 *Correlation Estimation*

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

THEOREM 1: *Under the assumption that the observed group responses are independent given the true group statuses, the covariance between Z_{jk} and $Z_{j'k}$, when written as a function of the correlation of the unknown individual responses, is*

$$\begin{aligned}Cov(Z_{jk},Z_{j'k}) = \Delta_{jj'}\Bigg[&\prod_{i=1}^{I_k}\left\{Corr(\tilde{Y}_{ijk},\tilde{Y}_{ij'k})\sqrt{Var(\tilde{Y}_{ijk})Var(\tilde{Y}_{ij'k})}+(1-\tilde{p}_{ijk})(1-\tilde{p}_{ij'k})\right\}-\\ &\prod_{i=1}^{I_k}(1-\tilde{p}_{ijk})(1-\tilde{p}_{ij'k})\Bigg]\end{aligned} \tag{5}$$

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

The proof is given in Web Appendix A. The importance of this theorem is that it provides a convenient way to obtain method of moments estimates for $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$. Suppose an estimate of the model given in Equation (1) is available so that we can then estimate θ_{jk}, denoted by $\hat{\theta}_{jk}$, through Equation (3). 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 $Cov(Z_{jk}, Z_{j'k})$ with $\hat{r}_{jk}\hat{r}_{j'k}$ in the left-hand side of Equation (5), we create one equation for each α_s $(s = 1, \ldots, S)$ and solve for α_s to obtain its estimate $\hat{\alpha}_s$. We argue in Web Appendix B that one unique solution $\hat{\alpha}_s$ can be found in each equation and that $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_S)$ is a consistent estimator of $\boldsymbol{\alpha}$ when $\boldsymbol{\beta}$ is known.

To illustrate, suppose there are possibly unequal working correlations among the individual disease response pairs, i.e., $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) = \alpha_{jj'}$ (we subscript the correlation parameter differently here to match the disease indices), so that $S = J(J-1)/2$. An estimate for $\alpha_{jj'}$ is obtained by solving

$$\sum_{k=1}^{K} \hat{r}_{jk}\hat{r}_{j'k} = \Delta_{jj'} \sum_{k=1}^{K} \left[\prod_{i=1}^{I_k} \left\{ \alpha_{jj'} \sqrt{\hat{\tilde{p}}_{ijk}(1-\hat{\tilde{p}}_{ijk})\hat{\tilde{p}}_{ij'k}(1-\hat{\tilde{p}}_{ij'k})} + (1-\hat{\tilde{p}}_{ijk})(1-\hat{\tilde{p}}_{ij'k}) \right\} - \prod_{i=1}^{I_k} (1-\hat{\tilde{p}}_{ijk})(1-\hat{\tilde{p}}_{ij'k}) \right] \quad (6)$$

for $\alpha_{jj'}$, where $\hat{\tilde{p}}_{ijk}$ is an estimate of $\tilde{p}_{ijk}$ that results from replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ in Equation (1). Alternatively, if one specifies an exchangeable correlation structure, i.e., $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) = \alpha$, only $S = 1$ equation needs to be solved. This equation is the same as in Equation (6), but with α replacing $\alpha_{jj'}$ and an additional summation $\sum_{j<j'}$ on both sides of the equality to sum over disease pairs.

Because $Cov(Z_{jk}, Z_{j'k})$ is a polynomial function of $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$ 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 higher order $(\geqslant 3)$ coefficients involving

$Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$ are almost always negligible (see Web Appendix C). As a result, it usually suffices to use the linear and quadratic terms to estimate $\boldsymbol{\alpha}$. For example, with an unstructured working correlation matrix, the linear and quadratic coefficients of $\alpha_{jj'}$ in Equation (6) are

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

and

$$\begin{aligned} \hat{d}_{jj',k} &= \Delta_{jj'} \left\{ \prod_{i=1}^{I_k} (1 - \hat{\tilde{p}}_{ijk})(1 - \hat{\tilde{p}}_{ij'k}) \right\} \times \\ &\quad \sum_{1 \leqslant i_1 < i_2 \leqslant I_k} \sqrt{\frac{\hat{\tilde{p}}_{i_1jk}\hat{\tilde{p}}_{i_1j'k}}{(1 - \hat{\tilde{p}}_{i_1jk})(1 - \hat{\tilde{p}}_{i_1j'k})}} \sqrt{\frac{\hat{\tilde{p}}_{i_2jk}\hat{\tilde{p}}_{i_2j'k}}{(1 - \hat{\tilde{p}}_{i_2jk})(1 - \hat{\tilde{p}}_{i_2j'k})}}, \end{aligned}$$

respectively. Web Appendix C provides specific details on how to obtain these coefficients. The estimate $\hat{\alpha}_{jj'}$ solves $\sum_{k=1}^{K} \hat{r}_{jk}\hat{r}_{j'k} = \sum_{k=1}^{K} \hat{c}_{jj',k}\hat{\alpha}_{jj'}$ using a first-order approximation or $\sum_{k=1}^{K} \hat{r}_{jk}\hat{r}_{j'k} = \sum_{k=1}^{K} (\hat{c}_{jj',k}\hat{\alpha}_{jj'} + \hat{d}_{jj',k}\hat{\alpha}_{jj'}^2)$ using a second-order approximation.

3.3 *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 hold in our setting too but with a small change to the form of $Cov(\hat{\boldsymbol{\beta}})$. Note that for each group k, the expectations $E(\tilde{Y}_{1jk}|Z_{jk}), \ldots, E(\tilde{Y}_{I_kjk}|Z_{jk})$ are all functions of Z_{jk}; thus, the $\Psi_{ik}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ 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, Equation 2.9) to incorporate this within group dependence (see Hardin and Hilbe, 2003, page 29). Specifically, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \widehat{Cov}(\hat{\boldsymbol{\beta}}) &= \left(\sum_k \sum_i \frac{\partial \Psi_{ik}^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \right)^{-1} \left\{ \sum_k \left(\sum_i \Psi_{ik}^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right) \left(\sum_i \Psi_{ik}^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right)' \right\} \times \\ &\quad \left(\sum_k \sum_i \frac{\partial \Psi_{ik}^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \right)^{-1} \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}, \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha}, \boldsymbol{D}_{ik}, \boldsymbol{V}_{ik}, \boldsymbol{\omega}_{ik},$ and $\tilde{\boldsymbol{p}}_{ik}$ are all functions of $\boldsymbol{\beta}$. Our simulation evidence in Section 4 shows that standard errors are estimated well by the corresponding entries in (7) and that resulting Wald confidence intervals confer nominal levels in realistic settings.

4. Simulation Evidence

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

$$\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik} \tag{8}$$

for $j = 1, 2$, where the between-disease correlation is $Corr(\tilde{Y}_{i1k}, \tilde{Y}_{i2k}) = \alpha$. 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 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 typical prevalence levels where group testing would be used. In Web Appendix D, we provide histograms of the true individual probabilities for one simulated data set.

We employ the following strategy to simulate the observed group responses Z_{jk} for $j = 1, 2$ and $k = 1, \ldots, K$. With individual probabilities from Equation (8) and a given value of α, we use the correlated binary data generation procedure of Emrich and Piedmonte (1991) to simulate the $(\tilde{Y}_{i1k}, \tilde{Y}_{i2k})$ responses, and 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 α and Equation

(8) for each of $B = 1000$ simulated data sets, where we estimate only one parameter, say β_2, for both β_{21} and β_{22} because these two parameters are equal. Note that in Section 5 we demonstrate two applications where it is sensible to share parameters across diseases (i.e., across the levels of j). Table 1 gives parameter estimates averaged over the simulated data sets for various combinations of α, K, and I_k ("Mean" row). The use of large sample sizes ($K \geqslant 500$) is motivated by our experience with the NPHL (see Section 5.1). In Table 1, one can see that the averaged estimates are all close to the true values. We also calculate the standard deviation (SD) for each regression parameter estimate across the simulated data sets and compare this to the corresponding averaged estimated standard error (SE) obtained from (7). The SE/SD ratio given in Table 1 approaches 1 as K increases, although SE is slightly underestimated for smaller K. Lastly, in Table 1, we give the estimated coverage probabilities 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.

[Table 1 about here.]

Given the previous work in group testing regression modeling, one may wonder how fitting J separate models compares to our multiple-disease model fit using the ES algorithm. The top portion of Table 2 compares variance estimates obtained through the ES algorithm (where one working correlation parameter α is estimated) to variance estimates obtained using the methods of Vansteelandt et al. (2000) which estimate separate models for $j = 1, 2$. Specifically, we calculate the relative efficiency as

$$RE(\hat{\beta}^{V}_{b,rj} \text{ to } \hat{\beta}^{ES}_{b,rj}) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{Var}(\hat{\beta}^{V}_{b,rj})}{\widehat{Var}(\hat{\beta}^{ES}_{b,rj})}, \tag{9}$$

where, for the b^{th} simulated data set, $\hat{\beta}^{ES}_{b,rj}$ denotes the r^{th} regression parameter estimate for the j^{th} disease using the ES algorithm and $\hat{\beta}^{V}_{b,rj}$ is the maximum likelihood estimate using the approach outlined in Vansteelandt et al. (2000). Note that we calculate the relative efficiency using $\widehat{Var}(\hat{\beta}^{ES}_{b,2})$ when $r = 2$ because the single shared parameter β_2 replaces $\beta_{21} = \beta_{22}$. For

relative efficiencies involving $\hat{\beta}_{b,2}^{ES}$, dramatic increases in efficiency can occur with levels at times greater than 2. In addition, even when parameters are not shared for $r = 1$, we still see important gains in efficiency (1.4% to 17.2%). To compare all regression estimators for each j, we include in Table 2 the relative efficiency as in Equation (9), but now involving $\widehat{Var}(\text{logit}(\hat{\hat{p}}_b))$ where $\hat{\hat{p}}$ denotes the estimated probability of disease positivity at the mean values of the two covariates in Equation (8). Again, we see the benefits of using the ES algorithm where the gains in efficiency range from 16.3% to 43.1%.

[Table 2 about here.]

The bottom portion of Table 2 provides the same comparisons as in the top portion, but with an independence working correlation structure (i.e., $\boldsymbol{R}(\boldsymbol{\alpha})$ is the identify matrix). Efficiency benefits from using the ES algorithm are still available; however, the benefits are generally not as large as those using the exchangeable structure. In other words, there are important gains in efficiency from estimating the within-subject correlation.

We have performed a number of additional simulations using different models, a larger number of diseases, smaller and larger prevalence levels, smaller sample sizes, and different levels of correlation among diseases. Details for some of these simulations are provided in Web Appendix D. For example, corresponding to Equation (8), we have also estimated β_{21} and β_{22} separately. While the gains in relative efficiency for this situation are less, they are still as large as 7.2%. In addition, we have used simulation settings somewhat similar to those observed in the prenatal infectious disease screening study described in Section 5.2. These simulations produced results comparable to those described above.

5. Applications

5.1 *NPHL*

Chlamydia and gonorrhea are the two most prevalent sexually transmitted diseases in the United States (Centers for Disease Control and Prevention, 2010). This is also true in Nebraska, where these diseases have been characterized as being at epidemic levels (Zagurski, 2006). As part of the Centers for Disease Control and Prevention funded IPP, the NPHL uses the GenProbe Aptima Combo 2 assay to test for chlamydia and gonorrhea simultaneously. Due to the high cost of individually testing about 25,000 people per year, the NPHL is interested in using group testing for screening. Other IPP participating laboratories, such as the State Hygienic Laboratory at the University of Iowa, already use group testing. Our goal is to fit models to 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 disease statuses are related to certain risk factors at a fraction of the cost when compared to testing subjects individually. 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, 2010).

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 $I_k = 5$ 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; e.g., see Morre et al. (2001). The NPHL's assay for female swabs has a sensitivity of 0.928 (0.966) for chlamydia (gonorrhea) and a specificity of 0.960 (0.980) for chlamydia (gonorrhea). We use these same levels here. In addition to the testing outcomes for both infections, the NPHL collects additional covariate information

on each individual. We use the following covariates in our models: age, race (represented by three indicator variables), symptoms, clinician observations (cervical friability, pelvic inflammatory disease, cervicitis), and risk history (multiple partners, new partner in the last 90 days, contact with someone who has a sexually transmitted disease). All covariates are dichotomous except for age.

Table 3 displays the results from fitting a first-order model using our methodology in Section 3 with a logit function as $f(\cdot)$ in Equation (1). The estimated value of α is 0.27, which is obtained using a second-order approximation. For comparison purposes, we also fit the same regression model using the individual observations with standard GEE methodology. When fitting the individual testing model, we assumed that $\eta_j = \delta_j = 1$. 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 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 negative.

[Table 3 about here.]

The parameter estimates given in Table 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 as many responses 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 those from individual testing. Using a 0.05 level of significance with the group testing models, Wald test p-values are less than 0.05 for the race*, symptoms, multiple partners*, and contact to a STD* covariates corresponding to gonorrhea, and the age*, race*,

symptoms, cervicitis, and contact to a STD covariates corresponding to chlamydia. Covariate effects listed with asterisks are significant when controlling the familywise error rate level at 0.05 with a Bonferroni adjustment. These results largely agree with those from fitting the individual testing model, although the individual testing analysis finds some additional estimates significant at the unadjusted 0.05 level.

Using our multiple-disease group testing model, it is possible to perform hypothesis tests of the form $H_0 : \beta_{r1} = \beta_{r2}$ versus $H_a : \beta_{r1} \neq \beta_{r2}$, for $r = 0, 1, \ldots, p$–1; i.e., we can test for a shared parameter between diseases. This type of test is helpful to determine if particular covariates, such as those involving sexual behavior, have a similar effect on different disease statuses. Note that this type of test can not 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 might be preferred to consider a more parsimonious model with a shared parameter across both diseases for these covariates. Sharing parameters across diseases also can lead to smaller standard errors for the corresponding estimators. When we estimate this model (see Web Appendix E), 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 estimate for cervicitis is now shared between the infections.

5.2 *Prenatal infectious disease screening*

Screening pregnant women for infectious diseases is important for public health purposes. However, in lesser developed countries, the scarcity of resources can make screening individuals too costly. Verstraeten et al. (1998) and Vansteelandt et al. (2000) examine a surveillance study in Kenya involving pregnant women monitoring disease prevalence in four

rural locations. For this study, women visited prenatal clinics to supply serum specimens, and these specimens were subsequently tested using both group and individual testing. The research showed that group testing provided similar estimates to those from individual testing when estimating the overall prevalence (Verstraeten et al., 1998) and covariate specific probabilities (Vansteelandt et al., 2000), while also providing a 62% reduction in costs.

In the data shared with us by Dr. Stijn Vansteelandt, there are 428 complete observations that include HIV, hepatitis B, and syphilis diagnoses for each individual. The overall prevalences for the infections are 0.082 for HIV, 0.075 for hepatitis B, and 0.026 for syphilis. In addition, covariate information on age, marital status (never been married, been married), and education level (1 = none, 2 = primary, 3 = secondary, and 4 = higher) are available on each individual. We therefore illustrate our multiple-disease regression methodology with the available data for all three infections. Unfortunately, the original group testing responses are no longer available, so we formed groups of size $I_k = 5$ ourselves by pooling individuals in the order as they appear in the data set. Also, sensitivity and specificity levels are not available for all diseases, so we use $\eta_j = \delta_j = 0.99$ for each disease.

Table 4 shows the parameter estimates from fitting a first-order model using the ES algorithm for the group responses and using the GEE methodology of Neuhaus (2002) for the individual responses. We use a logit function as $f(\cdot)$ in Equation (1), and we use a second-order approximation to estimate an unstructured working correlation matrix. Once again, we see general agreement between the group and individual testing model estimates. There are some minor differences (e.g., the syphilis intercept term), but nothing major given the corresponding estimated standard error levels. Similar to Section 5.1, these standard errors are approximately 0.8 to 3.0 times larger for the group testing model when compared to individual testing.

[Table 4 about here.]

We also include in Table 4 relevant Wald tests for this application. With the group testing model, we find marginal significance for the intercept, marital status, and education estimates. The individual testing model gives somewhat similar results, but with disagreement for marital status. We also perform Wald tests for the equality of regression parameters across the three diseases. Both models give strong evidence for differences among the education levels in how they are related to the disease statuses. Also, both methods give non-significant results for age and marital status.

6. Discussion

In this paper, we have generalized previous work in group testing regression to include multiple-disease data. When compared to the existing methodology, the proposed techniques allow for individual unobserved disease statuses to be modeled jointly while also incorporating testing error. We have also illustrated how to perform covariate-adjusted inferences across diseases and how our models can accommodate shared parameters. The website `www.chrisbilder.com/grouptesting/multiple` contains R functions that can be used to apply the methodology. We plan to include these functions within R's `binGroup` package (Bilder et al., 2010) in the near future.

Our proposed methodology could also be adapted to a single-disease longitudinal setting where individuals are pooled at each time point. This would involve simply letting the j subscript in our notation keep track of the time points for the i^{th} individual in the k^{th} group. One potential limitation with this extension is that the same individuals would need to be in the same groups at each time point, although this design does occur in related problems (see Malinovsky, Albert, and Schisterman (2012)). We have examined removing this design constraint, but we have found that it would be quite difficult to eliminate because Z_{jk} could be correlated with $I_k - 1$ other $Z_{j'k'}$ responses for $j \neq j'$and $k \neq k'$.

An alternative to our ES Algorithm fitting approach would be to include random effects in

Equation (1) to account for the correlation among disease responses within each individual. Only the work of Chen et al. (2009) has examined the use of random effects in a group testing regression context, and they do so for single-disease models. Using random effects would be much more difficult in the multiple-disease setting, because the likelihood function involves K different I_k dimensional integrals. Therefore, depending on the size of I_k, evaluating the likelihood function may be difficult or even intractable. Future research is needed to examine this potentially useful formulation.

An additional alternative approach would be to formulate a set of generalized estimating equations in terms of the observed group responses $\boldsymbol{Z}_k = (Z_{1k}, \ldots, Z_{Jk})'$ rather than in terms of the unobserved individual responses as we have done. This would be analogous to the approach taken by Vansteelandt et al. (2000) for single-disease models, and details of its implementation are given in Web Appendix F. While this alternative approach can provide similar estimates to those found in this paper, there are two main reasons not to use it. First, the working correlation structure must be specified in terms of the group responses, which is a very unnatural way to think about correlation in a group testing context (especially if different group sizes are used). Second, this approach can not be generalized to accommodate all group testing protocols that may be used in practice, such as when individuals are in more than one initial group, analogously to how the regression approach in Vansteelandt et al. (2000) can not be generalized within a single-disease setting.

On the other hand, when individuals are in more than one initial group and/or when retests are included, our ES algorithm approach can be generalized for these protocols. Similarly to how Xie (2001) does for single-disease models, one can reformulate the conditional expectations in Section 3.1 by taking into account the group testing protocol. When it is not possible to obtain a closed-form expression for these conditional expectations, one could use Gibbs sampling to approximate them. We also conjecture that incorporating information from

retests could sharpen the correlation estimates described in Section 3.2. However, because initial group responses are correlated with subsequent retest responses, formulating this extension could prove to be challenging.

Acknowledgements

This research is supported by Grant R01 AI067373 from the National Institutes of Health. The authors thank Dr. Peter Iwen, Dr. Steven Hinrichs, and Philip Medina for their consultation on chlamydia and gonorrhea screening by the NPHL. The authors also thank Dr. Stijn Vansteelandt and his colleagues for sharing the prenatal infectious disease data.

Supplementary Materials

Web appendices referenced in the manuscript are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

References

American Red Cross (2012). *Blood testing.* Retrieved April 7, 2012, from `http://www.redcrossblood.org/learn-about-blood/what-happens-donated-blood/blood-testing`.

Bilder, C. and Tebbs, J. (2009). Bias, efficiency, and agreement for group-testing regression models. *Journal of Statistical Computation and Simulation* **79**, 67-80.

Bilder, C., Tebbs, J., and Chen, P. (2010). Informative retesting. *Journal of the American Statistical Association* **105**, 942-955.

Bilder, C., Zhang, B., Schaarschmidt, F., and Tebbs, J. (2010). binGroup: a package for group testing. *The R Journal* **2**, 56-60.

Centers for Disease Control and Prevention (2010). *Sexually Transmitted Disease Surveillance 2009.* Atlanta: U.S. Department of Health and Human Services. Available at `http://www.cdc.gov/std/stats`.

Chen, P., Tebbs, J., and Bilder, C. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270-1278.

Chi, X., Lou, X., Yang, M., and Shu, Q. (2009). An optimal DNA pooling strategy for progressive fine mapping. *Genetica* **135**, 267-281.

Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association* **106**, 640-650.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13**, 48–65.

Emrich, L. and Piedmonte, M. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302-303.

Fahey, J., Ourisson, P., and Degnan, F. (2006). Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutrition Journal* **5**, 13.

Gaydos, C. (2005). Nucleic acid amplification tests for gonorrhea and chlamydia: practice and applications. *Infectious Disease Clinics of North America* **19**, 367-386.

Hardin, J. and Hilbe, J. (2003). *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Huang, X. (2009). Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics* **65**, 361-368.

Hughes-Oliver, J. and Rosenberger, W. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika* **87**, 315-327.

Malinovsky, Y., Albert, P., and Schisterman, E. (2012). Pooling designs for outcomes under a Gaussian random effects model. *Biometrics* **68**, 45-52.

Morre, S., Dijk, R., Meijer, C., Brule, A., Kjaer, S., and Munk, C. (2001). Pooling cervical

swabs for detection of chlamydia trachomatis by PCR: sensitivity, dilution, inhibition, and cost-saving aspects. *Journal of Clinical Microbiology* **39**, 2375-2376.

Munoz-Zanzi, C., Thurmond, M., Hietala, S., and Johnson, W. (2006). Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. *Preventative Veterinary Medicine* **74**, 309-322.

Neuhaus, J. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* **58**, 675-683.

Stramer, S., Glynn, S., Kleinman, S., Strong, D., Caglioti, S., Wright, D., Dodd, R., and Busch, M. (2004). Detection of HIV-1 and HCV infections among antibody-negative blood donors by nucleic acid–amplification testing. *New England Journal of Medicine* **351**, 760-768.

Tebbs, J. and Bilder, C. (2004). Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics* **9**, 75-90.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126-1133.

Verstraeten, T., Farah, B., Duchateau, L., and Matu, R. (1998). Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district. *Tropical Medicine and International Health* **3**, 747-750.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine* **20**, 1957-1969.

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 1

Simulation results using the ES algorithm to estimate the model given in Equation (8) *with* $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1,$ *and* $\beta_2 = 0.1$. *The estimated parameters and their standard errors are averaged over 1000 simulated data sets. The estimated coverage probabilities are for 95% Wald confidence intervals involving the corresponding regression parameter.*

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_2$	$\hat{\alpha}$
0.6	1000	5	Mean	-5.99	-7.03	-0.03	1.00	0.10	0.61
			SE/SD	0.96	0.96	0.98	0.95	0.96	—
			Coverage	0.95	0.94	0.95	0.95	0.95	—
	500	10	Mean	-6.14	-7.20	0.00	1.07	0.10	0.61
			SE/SD	0.99	0.97	0.94	0.93	0.95	—
			Coverage	0.94	0.94	0.95	0.95	0.94	—
0.2	1000	5	Mean	-6.02	-7.03	0.03	1.02	0.10	0.20
			SE/SD	0.95	0.95	0.94	0.95	0.96	—
			Coverage	0.94	0.95	0.95	0.95	0.94	—
	500	10	Mean	-6.12	-7.21	0.01	1.13	0.10	0.21
			SE/SD	0.97	0.98	0.96	0.96	0.98	—
			Coverage	0.94	0.94	0.95	0.95	0.95	—
0.6	2000	5	Mean	-6.00	-7.02	0.00	1.02	0.10	0.60
			SE/SD	0.98	1.00	0.95	0.99	1.00	—
			Coverage	0.94	0.94	0.94	0.96	0.95	—
	1000	10	Mean	-6.01	-7.04	0.04	1.06	0.10	0.60
			SE/SD	0.99	1.00	0.96	0.98	0.99	—
			Coverage	0.95	0.96	0.95	0.95	0.95	—
0.2	2000	5	Mean	-6.02	-7.05	0.01	1.04	0.10	0.20
			SE/SD	0.97	0.96	1.01	1.00	0.96	—
			Coverage	0.94	0.94	0.96	0.96	0.94	—
	1000	10	Mean	-6.05	-7.06	0.03	1.03	0.10	0.20
			SE/SD	0.97	1.00	0.96	0.99	0.97	—
			Coverage	0.95	0.94	0.95	0.95	0.95	—

Table 2
Relative efficiency of the parameter estimates for the model in Equation (8).

ES algorithm with exchangeable correlation structure

α	K	I_k	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\text{logit}(\hat{\tilde{p}})$ $j=1$	$j=2$
0.6	1000	5	1.249	1.669	1.085	1.067	1.290	2.229	1.211	1.279
	500	10	1.287	1.737	1.113	1.172	1.335	2.358	1.287	1.431
0.2	1000	5	1.409	1.828	1.049	1.088	1.573	2.598	1.174	1.326
	500	10	1.469	1.897	1.079	1.136	1.718	2.817	1.264	1.404
0.6	2000	5	1.197	1.575	1.050	1.014	1.237	1.984	1.163	1.224
	1000	10	1.242	1.584	1.061	1.074	1.312	1.999	1.218	1.264
0.2	2000	5	1.373	1.733	1.016	1.032	1.521	2.411	1.173	1.275
	1000	10	1.462	1.758	1.038	1.070	1.655	2.455	1.241	1.340

ES algorithm with independence correlation structure

α	K	I_k	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\text{logit}(\hat{\tilde{p}})$ $j=1$	$j=2$
0.6	1000	5	1.198	1.641	1.028	1.065	1.267	2.149	1.080	1.272
	500	10	1.240	1.665	1.088	1.145	1.318	2.236	1.146	1.376
0.2	1000	5	1.392	1.805	1.038	1.078	1.558	2.559	1.159	1.318
	500	10	1.460	1.862	1.077	1.128	1.715	2.761	1.252	1.382
0.6	2000	5	1.149	1.552	0.993	1.013	1.211	1.932	1.072	1.223
	1000	10	1.223	1.566	1.037	1.066	1.308	1.951	1.126	1.256
0.2	2000	5	1.357	1.718	1.004	1.025	1.507	2.383	1.157	1.270
	1000	10	1.459	1.753	1.037	1.069	1.654	2.445	1.235	1.339

Table 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. Tests for significance of the intercept parameters are not of interest for this example, so we exclude these p-values. Note that we perform one joint test for each disease when evaluating race.

		ES algorithm		GEE	
Disease	Term	Estimate(SE)	P-value	Estimate(SE)	P-value
Gonorrhea	Intercept	-5.722(0.605)	NA	-4.553(0.327)	NA
	Age	-0.031(0.021)	0.1451	-0.040(0.013)	0.0018
	Race level #1	2.020(0.359)	<0.0001	1.319(0.173)	<0.0001
	Race level #2	0.771(1.080)		0.715(0.336)	
	Race level #3	0.782(0.857)		-0.113(0.425)	
	Symptoms	1.092(0.384)	0.0045	0.930(0.175)	<0.0001
	Cervical friability	-0.194(0.648)	0.7645	0.325(0.312)	0.2960
	Pelvic inflam. disease	0.283(0.963)	0.7685	1.158(0.524)	0.0272
	Cervicitis	0.293(0.349)	0.4010	0.550(0.200)	0.0060
	Multiple partners	1.167(0.311)	0.0002	1.046(0.171)	<0.0001
	New partner	0.292(0.332)	0.3804	-0.086(0.186)	0.6422
	Contact to a STD	1.381(0.286)	<0.0001	1.170(0.181)	<0.0001
Chlamydia	Intercept	-0.520(0.419)	NA	-0.976(0.147)	NA
	Age	-0.113(0.019)	<0.0001	-0.088(0.007)	<0.0001
	Race level #1	0.591(0.120)	<0.0001	0.392(0.096)	<0.0001
	Race level #2	1.062(0.243)		0.691(0.136)	
	Race level #3	0.036(0.401)		0.057(0.151)	
	Symptoms	0.385(0.175)	0.0280	0.287(0.082)	0.0005
	Cervical friability	0.309(0.305)	0.3104	0.056(0.170)	0.7420
	Pelvic inflam. disease	0.788(0.627)	0.2089	0.400(0.387)	0.3016
	Cervicitis	0.534(0.199)	0.0074	0.591(0.107)	<0.0001
	Multiple partners	0.279(0.221)	0.2059	0.468(0.100)	<0.0001
	New partner	0.064(0.197)	0.7435	-0.044(0.092)	0.6368
	Contact to a STD	0.591(0.212)	0.0053	0.935(0.101)	<0.0001

Table 4

Parameter estimates and estimated standard errors (in parentheses) for the prenatal infectious disease screening data. Marital status is represented by an indicator variable (1 = never married; 0 = been married). The GEE columns correspond to a model fit to the individual testing responses using the methodology of Neuhaus (2002). The "Overall test" column contains p-values for the Wald test $H_0 : \beta_{r1} = \beta_{r2} = \beta_{r3} = 0$ *versus* H_a *: At least one* β_{rj} *not equal to 0, where 1 = syphilis, 2 = hepatitis B, 3 = HIV, and r denotes the covariate of interest. The "Across test" column contains p-values for the Wald test* $H_0 : \beta_{r1} = \beta_{r2} = \beta_{r3}$ *versus* H_a *: At least one* β_{rj} *unequal.*

Method	Term	Disease	Estimate(SE)	Overall test	Across test
ES algorithm	Intercept	Syphilis	-0.749(2.019)	0.070	0.420
		Hepatitis B	-2.115(2.061)		
		HIV	-4.531(1.913)		
	Age	Syphilis	0.014(0.061)	0.961	0.939
		Hepatitis B	-0.005(0.075)		
		HIV	0.029(0.060)		
	Marital Status	Syphilis	-0.799(2.835)	0.075	0.191
		Hepatitis B	1.827(0.740)		
		HIV	-0.663(1.498)		
	Education	Syphilis	-1.922(1.059)	0.017	0.008
		Hepatitis B	-0.416(0.408)		
		HIV	0.663(0.351)		
GEE	Intercept	Syphilis	0.279(1.638)	$<$0.001	0.031
		Hepatitis B	-1.782(1.118)		
		HIV	-4.026(0.844)		
	Age	Syphilis	-0.094(0.076)	0.535	0.516
		Hepatitis B	-0.025(0.033)		
		HIV	-0.001(0.030)		
	Marital Status	Syphilis	-0.593(1.749)	0.376	0.784
		Hepatitis B	0.646(0.534)		
		HIV	0.702(0.495)		
	Education	Syphilis	-1.145(0.837)	0.002	0.009
		Hepatitis B	-0.170(0.279)		
		HIV	0.592(0.172)		