Published in final edited form as:

Stat Med. 2013 December 10; 32(28): 4954–4966. doi:10.1002/sim.5858.

Regression analysis for multiple-disease group testing data

Boan Zhang^a, Christopher R. Bilder^a, and Joshua M. Tebbs^{b,*}

^aDepartment of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

^bDepartment of Statistics, University of South Carolina, Columbia, SC 29208, USA

Abstract

Group testing, where individual specimens are composited into groups to test for the presence of a disease (or other binary characteristic), 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 available, but inferences are needed at the individual level. A further methodological challenge arises when individuals are tested in groups for multiple diseases simultaneously, because unobserved individual disease statuses are likely correlated. In this paper, we propose new 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.

Keywords

correlated binary data; expectation-solution algorithm; generalized estimating equations; Infertility Prevention Project; pooled testing; specimen pooling

1. Introduction

Medical 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. This has led to the adoption of group testing in a number of infectious disease applications, including blood donation screening by the American Red Cross [1], opportunistic chlamydia and gonorrhea testing in medical clinics [2], and detecting influenza viruses in humans [3]. Group testing has also proven to be beneficial in other areas including drug discovery [4], genetics [5], animal ecology [6], and food contamination testing [7].

Statistical research in group testing has traditionally focused on estimating the overall disease prevalence. More recently, this research has shifted towards incorporating covariate information to produce individual-specific estimates in a regression context; Vansteelandt *et al.* [8] and Xie [9] are commonly regarded as the seminal papers in this area. Vansteelandt *et al.* [8] provides a generalized linear model approach that uses only the initial group responses for estimation. The approach taken by Xie [9] 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.* [8] and Xie [9]. Bilder and Tebbs [10] provide a thorough comparison of individual and group testing regression model estimates, Chen *et al.* [11] examine mixed-effects models, and Delaigle and Meister [12] and Delaigle and Hall [13] propose nonparametric modeling approaches. Group testing regression methods also have been used to detect model misspecification with individual response data, as shown by Huang [14].

When viewed collectively, research in group testing regression modeling has had one common theme; 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 [1, 15]. 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 [16] is the only work that has addressed multiple infections in the group testing literature, and they do so by estimating overall prevalences under the assumption that diagnostic tests are perfect.

In this 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.* [8] and Xie [9] to model multiple-disease statuses and as a generalization of Hughes-Oliver and Rosenberger [16] 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 [17] 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 true individual responses by using the observed (possibly misclassified) 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. Section 6 summarizes this work and suggests future areas of research.

2. Notation, assumptions, and model

Let $\tilde{Y}_{ijk}=1$ (0) if individual i in group k is truly positive (negative) for disease j , for $i = 1, \dots$

$I_k, j = 1, \dots, J$, and $k = 1, \dots, K$. We assume that $\tilde{\mathbf{Y}}_{ik} = (\tilde{Y}_{i1k}, \dots, \tilde{Y}_{iJk})'$ are independent random vectors across i and k but allow for $\tilde{Y}_{i1k}, \dots, \tilde{Y}_{iJk}$ to be 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 [16], $Z_{jk} = 1$ 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 for infectious diseases 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 do not depend on pool sizes or covariates; these assumptions are analogous to those made by Vansteelandt *et al.* [8] and Xie [9] for single-disease group testing models and by Neuhaus [18] for individual testing models.

With covariates $\mathbf{x}_{ik} = (x_{1ik}, \dots, x_{m-1,ik})'$ collected on each individual, our goal is to estimate $P(\tilde{Y}_{ijk}=1 | \mathbf{x}_{ik}) \equiv \tilde{p}_{ijk}$ when only the observed group responses Z_{jk} are available, similar to Vansteelandt *et al.* [8] with single-disease models. In all subsequent expectations written in this paper, we condition on the full set of covariates $\mathbf{x}_{ik}$ as we did for $\tilde{p}_{ijk}$, but we suppress this specification for notational simplicity. We consider models of the form

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

where $f(\cdot)$ is a known monotonic, differentiable function and β_{rj} ($r = 0, \dots, m-1; j = 1, \dots, J$) is a regression parameter. Using a joint model, as in Equation (1), 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 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. Expectation-Solution algorithm

We use the ES Algorithm to estimate the parameters in Equation (1). The ES algorithm, introduced by Elashoff and Ryan [17], is a generalization of the expectation-maximization (EM) algorithm given by Dempster *et al.* [19]. 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 [17] 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. Our application of the ES algorithm requires additional work to estimate the correlation among the unobserved individual responses, as shown in Section 3.2.

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 (1). Let $\mathbf{R}(\alpha)$ denote the $J \times J$ working correlation matrix for the true individual responses that is dependent on an $S \times 1$ vector α . Define $cov(\tilde{\mathbf{Y}}_{ik}) = \mathbf{V}_{ik} = \mathbf{B}_{ik}^{1/2} \mathbf{R}(\alpha) \mathbf{B}_{ik}^{1/2}$, where $\mathbf{B}_{ik} = \text{diag}(\tilde{p}_{ijk}(1 - \tilde{p}_{ijk}))$. The estimating equations are

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

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

Because the individual responses $\tilde{Y}_{ijk}$ are not observed in group testing, we can not use standard GEE methodology as stated above. However, analogous to the use of the EM algorithm described in Xie [9] for a single disease, we can replace the individual responses in Equation (2) by their expected values, conditional on the group responses $\mathbf{Z} = (Z_{11}, \dots, Z_{JK})'$. Because a conditional expectation involving $\tilde{Y}_{ijk}$ depends only on its corresponding group response, it suffices to calculate $E(\tilde{Y}_{ijk} | Z_{jk}=1) = \eta_j \tilde{p}_{ijk} / \theta_{jk}$ and $E(\tilde{Y}_{ijk} | Z_{jk}=0) = (1 - \eta_j) \tilde{p}_{ijk} / (1 - \theta_{jk})$, where

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

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

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

where $\omega_{ik} = (E(\tilde{Y}_{i1k}|z_{1k}), \dots, E(\tilde{Y}_{iJk}|z_{Jk}))'$ and $\Psi_{ik}^{obs}(\beta, \alpha) = \mathbf{D}_{ik}' \mathbf{V}_{ik}^{-1} (\omega_{ik} - \tilde{\mathbf{p}}_{ik})$. The ES Algorithm successively estimates α and solves Equation (4) for β in an iterative manner to obtain parameter estimates. The initial estimate of β can be found by estimating separate models for each disease with the methodology in Xie [9]. Note that the expectations

$E(\tilde{Y}_{ijk}|z_{jk})$ are updated at each iteration to correspond to the current estimate of β . Estimating α at each iteration is not straightforward, so we discuss it thoroughly in the next subsection. The final iterative solution to Equation (4) at convergence is the estimate of β , which we denote by $\hat{\beta}$. Consistency of $\hat{\beta}$ follows immediately from the results given in Section 2.1 of Elashoff and Ryan [17].

3.2. Correlation estimation

To estimate α , we need to first identify the relationship between $\text{cov}(Z_{jk}, Z_{j'k})$, which we can estimate from the observed group responses, and $\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$, 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 $Z_{j'k}$, when written as a function of the correlation of the unknown individual responses, is

$$\begin{aligned} & \text{cov}(Z_{jk}, Z_{j'k}) \\ &= \Delta_{jj'} \left[\prod_{i=1}^{I_k} \left\{ \text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) \sqrt{\text{var}(\tilde{Y}_{ijk}) \text{var}(\tilde{Y}_{ij'k})} + (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} - \prod_{i=1}^{I_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right] \quad (5) \end{aligned}$$

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

The proof of Theorem 1 is given in the Web-based Supporting Materials. The importance of Theorem 1 is that it provides a convenient way to obtain method of moments estimates for $\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$. 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 $\text{cov}(Z_{jk}, Z_{j'k})$ with $\hat{r}_{jk}\hat{r}_{j'k}$ in the left-hand side of Equation (5), we create one equation for each element of α , say, α_s ($s = 1, \dots, S$), and solve for α_s to obtain its estimate $\hat{\alpha}_s$. We argue in the Web-based Supporting Materials that one unique solution $\hat{\alpha}_s$ can be found in each equation and that $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_S)'$ is a consistent estimator of α .

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

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

for $\alpha_{jj'}$, where $\hat{p}_{ijk}$ is an estimate of $\tilde{p}_{ijk}$ that results from replacing β with $\hat{\beta}$ in Equation (1). Alternatively, if one specifies an exchangeable correlation structure, i.e.,

$\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) = \alpha$, 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 $\text{cov}(Z_{jk}, Z_{j'k})$ is a polynomial function of $\text{corr}(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k})$ of degree I_k , as shown in Theorem 1, 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 (3)

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

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

and

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

respectively. 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. More details on these approximations, including their derivations and accuracy, are available in the Web-based Supporting Materials.

3.3. Variance estimation

Elashoff and Ryan [17] 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 $\text{cov}(\hat{\beta})$. Note that for each group k , the expectations

$E(\tilde{Y}_{1jk}|Z_{jk}), \dots, E(\tilde{Y}_{i_kjk}|Z_{jk})$ are all functions of Z_{jk} ; 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 [17] to incorporate this within group dependence. Specifically, the covariance matrix of $\hat{\beta}$ is

$$\text{cov}(\hat{\beta}) = \left(\sum_k \sum_i \frac{\partial \Psi_{ik}^{obs}(\beta, \alpha)}{\partial \beta} \right)^{-1} \left\{ \sum_k \left(\sum_i \Psi_{ik}^{obs}(\beta, \alpha) \right) \left(\sum_i \Psi_{ik}^{obs}(\beta, \alpha) \right)' \right\} \left(\sum_k \sum_i \frac{\partial \Psi_{ik}^{obs}(\beta, \alpha)}{\partial \beta} \right)^{-1}, \quad (7)$$

where α , $\mathbf{D}_{ik}$, $\mathbf{V}_{ik}$, ω_{ik} , and $\tilde{\mathbf{p}}_{ik}$ are all functions of β . An estimate of this covariance matrix, which we denote by $\hat{\text{cov}}(\hat{\beta})$ arises from evaluating Equation (7) at $\hat{\alpha}$ and $\hat{\beta}$. Our simulation evidence in Section 4 shows that standard errors are estimated well by the corresponding entries in $\hat{\text{cov}}(\hat{\beta})$ and that resulting Wald confidence intervals confer nominal c 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:

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

for $j = 1, 2$, where the between-disease correlation is $\text{corr}(\tilde{Y}_{i1k}, \tilde{Y}_{i2k}) = \alpha$. 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.

We employ the following strategy to simulate the observed group responses Z_{jk} for $j = 1, 2$ and $k = 1, \dots, K$. With individual probabilities from Equation (8) and a given value of α , we use the correlated binary data generation procedure of Emrich and Piedmonte [20] 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 $\beta_{21} = \beta_{22}$ because these two parameters are assumed to be equal. This is motivated by our analysis of real data in Section 5 where we consider two different data sets; in each one,

the hypothesis of sharing parameters across diseases for a certain covariate (i.e., across the levels of j) is not rejected. 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 is motivated by our experience with the NPHL (see Section 5.1) and because group testing is frequently used in high volume clinical specimen screening situations. 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 Equation (7). The SE/SD ratio given in Table 1 approaches 1 as K increases, but the SE can be underestimated for smaller K especially for larger I_k . This behavior was also observed in Bilder and Tebbs [10] with single-disease group testing regression models. Lastly, in Table 1, we give the estimated coverage probabilities of nominal 95% Wald confidence intervals for each regression parameter. These levels are all between 0.93 and 0.96 (except for the $I_k = 10$ and $K = 250$ cases), which indicates the intervals are generally performing as expected.

Given the previous work in group testing regression modeling, one might legitimately wonder how fitting J separate models would compare to our joint multiple-disease model fit using the ES algorithm. Table 2 compares variance estimates obtained through the ES algorithm, where one working correlation parameter is estimated, to variance estimates obtained using the maximum likelihood methods of Vansteelandt *et al.* [8] which estimate separate models for $j = 1, 2$. Specifically, we calculate the relative efficiency as

$$RE\left(\hat{\beta}_{b,rj}^{ML} \text{ to } \hat{\beta}_{b,rj}^{ES}\right) = \frac{1}{B} \sum_{b=1}^B \frac{v\hat{a}r\left(\hat{\beta}_{b,rj}^{ML}\right)}{v\hat{a}r\left(\hat{\beta}_{b,rj}^{ES}\right)}, \quad (9)$$

where, for the b th simulated data set, $\hat{\beta}_{b,rj}^{ES}$ denotes the r th regression parameter estimate for the j th disease using the ES algorithm and $\hat{\beta}_{b,rj}^{ML}$ is the maximum likelihood estimate using the approach outlined in Vansteelandt *et al.* [8]. Note that we calculate the relative efficiency using $v\hat{a}r\left(\hat{\beta}_{b,2}^{ES}\right)$ when $r = 2$ because the single parameter β_2 replaces $\beta_{21} = \beta_{22}$.

For relative efficiencies involving $\hat{\beta}_{b,2}^{ES}$, dramatic increases in efficiency are seen in Table 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. For example, a 11.3% gain is achieved when $\alpha = 0.6$, $K = 500$, $I_k = 10$, and $j = 1$. To compare all regression estimators for each j , we also include in Table 2 the relative efficiency as in Equation (9), but now involving

$v\hat{a}r\left(\text{logit}\left(\hat{p}_b\right)\right)$ where $\hat{p}$ denotes the estimated probability of disease positivity at the mean values of the two covariates in Equation (8). Again, we see the benefits of joint modeling with gains in efficiency. For example, a 43.1% gain is achieved when $\alpha = 0.6$, $K = 500$, and $I_k = 10$.

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 many of these simulations are provided in the Web-based Supporting Materials. For example, corresponding to Equation (8), we have also performed simulations where β_{21} and β_{22} are estimated separately. It is not surprising that the relative efficiency gains in this situation are smaller, but they are still as large as 11%. In addition, we have used simulation settings 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. The one notable difference is that the SE/SD ratios did not show signs of underestimation despite the smaller sample sizes, which is likely because the mean prevalences were larger.

5. Applications

5.1. Nebraska IPP data

Chlamydia and gonorrhea are the two most common sexually transmitted diseases in the United States [21]. This is also true in Nebraska, where both diseases have been characterized as being at epidemic levels [22]. As part of the IPP, the NPHL uses the GenProbe Aptima Combo 2 assay to test higher-risk individuals for chlamydia and gonorrhea simultaneously. Due to the high cost of individual testing for 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 and the Idaho Bureau of Laboratories [23], 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 [24].

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 [25, 26, 27]. 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 the methodology in Section 3 with a logit function as $f(\cdot)$ in Equation (1); model fits for other group sizes are given in the Web-based Supporting Materials. The estimated value of α is $\hat{\alpha}=0.27$, which is obtained using a second-order approximation as described in Section 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 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 [18], 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 actually negative.

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.* [8] and Bilder and Tebbs [10] 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 model, it is straightforward to perform hypothesis tests of the form $H_0: \beta_{r1} = \beta_{r2}$ versus $H_1: \beta_{r1} \neq \beta_{r2}$, for $r = 0, 1, \dots, m - 1$; i.e., to test for a shared parameter between diseases. This type of test might be of interest if one goal is to determine if particular covariates, such as those involving sexual behavior, have a similar effect on different disease statuses. 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. When we fit this reduced model (see the Web-based Supporting Materials), we found that Wald test p-values were less than 0.05 for the same covariates as before. The only difference was that the significant estimate for cervicitis was shared between the infections.

5.2. Prenatal infectious disease screening data

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.* [28] and Vansteelandt *et al.* [8] summarize 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 [28] and covariate specific probabilities [8], 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. Many other group sizes are used in practice for these diseases, but we choose a small group size here to ensure that an adequate number of group responses are available for model fitting. Note that sensitivity and specificity levels are not available for all diseases, so we use values of $\eta_j = \delta_j = 0.99$ for each disease. These levels are reasonable given the information available on assays used in application [29, 30, 31].

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 [18] for the individual responses. We use a logit function for $f(\cdot)$ in Equation (1), and we use a second-order approximation to estimate an unstructured working correlation matrix. The working correlation estimates for the group testing model are 0.3354 for syphilis and hepatitis B, 0.1973 for syphilis and HIV, and 0.0332 for hepatitis B and HIV. 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 errors. 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.

We also include in Table 4 the 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 models give non-significant results for age and marital status.

6. Discussion

We have generalized previous work in group testing regression to include multiple-disease data. When compared to existing methods, our proposed techniques allow for unobserved individual disease statuses to be modeled jointly while also incorporating testing error. We have also illustrated how to estimate the correlation between unobserved disease statuses and how to perform covariate-adjusted inferences across diseases. The web site

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 [32] in the near future.

It is interesting to note that the framework proposed herein could be easily 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 has been proposed in related problems where pooling is used [33].

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 Chen *et al.* [11] have examined 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 directly may be difficult or even intractable. Future research is needed to examine this potentially useful formulation.

A second alternative approach would be to formulate a set of generalized estimating equations in terms of the observed group responses $\mathbf{Z}_k = (Z_{1k}, \dots, Z_{Jk})'$ 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.* [8] for single-disease models. 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 would have to be specified in terms of the latent 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 [34], analogously to how the regression approach in Vansteelandt *et al.* [8] can not be immediately 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 [9] does for single-disease models, one can reformulate the conditional expectations in Section 3.1 by taking into account the group testing protocol used. 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 precisely could prove to be challenging.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was 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. Suggestions given by the Associate Editor and a referee helped to improve this manuscript for which we are very appreciative.

References

- [1]. American Red Cross. Blood testing. Available at <http://www.redcrossblood.org/learn-about-blood/what-happens-donated-blood/blood-testing>. Retrieved April 1, 2013
- [2]. Gaydos C. Nucleic acid amplification tests for gonorrhea and chlamydia: Practice and applications. *Infectious Disease Clinics of North America*. 2005; 19:367–386. DOI: 10.1016/j.bbr.2011.03.031. [PubMed: 15963877]
- [3]. Van T, Miller J, Warshauer D, Reisdorf E, Jerrigan D, Humes R, Shult P. Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology*. 2012; 50:891–896. DOI: 10.1128/JCM.05631-11. [PubMed: 22205820]
- [4]. Remlinger K, Hughes-Oliver J, Young S, Lam R. Statistical design of pools using optimal coverage and minimal collision. *Technometrics*. 2006; 48:133–143. DOI: 10.1198/004017005000000481.
- [5]. Chi X, Lou X, Yang M, Shu Q. An optimal DNA pooling strategy for progressive fine mapping. *Genetica*. 2009; 135:267–281. DOI: 10.1007/s10709-008-9275-5. [PubMed: 18506582]
- [6]. Dhand N, Johnson W, Toribio J. A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*. 2010; 15:452–473. DOI: 10.1007/s13253-010-0032-8.
- [7]. Fahey J, Ourisson P, Degnan F. Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutrition Journal*. 2006; 5:13. DOI: 10.1186/1475-2891-5-13. [PubMed: 16630354]
- [8]. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. DOI: 10.1111/j.0006-341X.2000.01126.x. [PubMed: 11129470]
- [9]. Xie M. Regression analysis of group testing samples. *Statistics in Medicine*. 2001; 20:1957–1969. DOI: 10.1002/sim.817. [PubMed: 11427952]
- [10]. Bilder C, Tebbs J. Bias, efficiency, and agreement for group-testing regression models. *Journal of Statistical Computation and Simulation*. 2009; 79:67–80. DOI:10.1080/00949650701608990. [PubMed: 20046956]
- [11]. Chen P, Tebbs J, Bilder C. Group testing regression models with fixed and random effects. *Biometrics*. 2009; 65:1270–1278. DOI: 10.1111/j.1541-0420.2008.01183.x. [PubMed: 19210734]
- [12]. Delaigle A, Meister A. Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*. 2011; 106:640–650. DOI: 10.1198/jasa.2011.tm10520.
- [13]. Delaigle A, Hall P. Nonparametric regression with homogeneous group testing data. *Annals of Statistics*. 2012; 40:131–158. DOI: 10.1214/11-AOS952.
- [14]. Huang X. Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics*. 2009; 65:361–368. DOI: 10.1111/j.1541-0420.2008.01103.x. [PubMed: 18759837]
- [15]. Stramer S, Glynn S, Kleinman S, Strong D, Caglioti S, Wright D, Dodd R, Busch M. Detection of HIV-1 and HCV infections among antibody-negative blood donors by nucleic acid-amplification testing. *New England Journal of Medicine*. 2004; 351:760–768. DOI: 10.1056/NEJMoa040085. [PubMed: 15317889]
- [16]. Hughes-Oliver J, Rosenberger W. Efficient estimation of the prevalence of multiple rare traits. *Biometrika*. 2000; 87:315–327. DOI: 10.1093/biomet/87.2.315.
- [17]. Elashoff M, Ryan L. An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*. 2004; 13:48–65. DOI: 10.1198/1061860043092.

- [18]. Neuhaus J. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*. 2002; 58:675–683. DOI: 10.1111/j.0006-341X.2002.00675.x. [PubMed: 12230004]
- [19]. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–22. DOI: 10.2307/2984875.
- [20]. Emrich L, Piedmonte M. A method for generating high-dimensional multivariate binary variates. *American Statistician*. 1991; 45:302–303. DOI: 10.1080/00031305.1991.10475828.
- [21]. Centers for Disease Control and Prevention. Sexually Transmitted Disease Surveillance 2009. U.S. Department of Health and Human Services; Atlanta: Available at <http://www.cdc.gov/std/stats09/default.htm>. Retrieved April 1, 2013
- [22]. Zagurski K. 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*. Feb 2.2006 :08B.
- [23]. Lewis J, Lockary V, Kobic S. Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases*. 2012; 39:46–48. DOI: 10.1097/OLQ.0b013e318231cd4a. [PubMed: 22183846]
- [24]. Bilder C, Tebbs J, Chen P. Informative retesting. *Journal of the American Statistical Association*. 2010; 105:942–955. DOI: 10.1198/jasa.2010.ap09231. [PubMed: 21113353]
- [25]. Morre S, Dijk R, Meijer C, Brule A, Kjaer S, Munk C. Pooling cervical swabs for detection of chlamydia trachomatis by PCR: sensitivity, dilution, inhibition, and cost-saving aspects. *Journal of Clinical Microbiology*. 2001; 39:2375–2376. DOI: 10.1128/JCM. [PubMed: 11414248]
- [26]. Rours G, Verkooyen R, Willemse H, van der Zwaan E, van Belkum A, de Groot R, Verbrugh H, Ossewaarde J. Use of pooled urine samples and automated DNA isolation to achieve improved sensitivity and cost-effectiveness of large-scale testing for chlamydia trachomatis in pregnant women. *Journal of Clinical Microbiology*. 2005; 43:4684–4690. DOI: 10.1128/JCM.43.9.4684-4690.2005. [PubMed: 16145127]
- [27]. Clark A, Steece R, Crouse K, Campbell J, Zanto S, Kartchner D, Mottice S, Pettit D. Multisite pooling study using ligase chain reaction in screening for genital chlamydia trachomatis infections. *Sexually Transmitted Diseases*. 2001; 28:565–568. DOI: 10.1097/00007435-200110000-00002. [PubMed: 11689754]
- [28]. Verstraeten T, Farah B, Duchateau L, Matu R. Pooling sera to reduce the cost of HIV surveillance: A feasibility study in a rural Kenyan district. *Tropical Medicine and International Health*. 1998; 3:747–750. DOI: 10.1046/j.1365-3156.1998.00293.x. [PubMed: 9754671]
- [29]. Chou R, Huffman L, Fu R, Smits A, Korthuis P. Screening for HIV: A review of the evidence for the U.S. preventive services task force. *Annals of Internal Medicine*. 2005; 143:55–73. [PubMed: 15998755]
- [30]. World Health Organization. Hepatitis B Surface Antigen Assays: Operational Characteristics, Report 2. Available at <http://apps.who.int/iris/handle/10665/43031>. Retrieved April 1, 2013
- [31]. U.S. Preventive Services Task Force. Screening for syphilis infection: Recommendation statement. *Annals of Family Medicine*. 2004; 2:362–365. DOI: 10.1370/afm.215. [PubMed: 15335137]
- [32]. Bilder C, Zhang B, Schaarschmidt F, Tebbs J. binGroup: A package for group testing. *R Journal*. 2010; 2:56–60. [PubMed: 21833347]
- [33]. Malinovsky Y, Albert P, Schisterman E. Pooling designs for outcomes under a Gaussian random effects model. *Biometrics*. 2012; 68:45–52. DOI: 10.1111/j.1541-0420.2011.01673.x. [PubMed: 21981372]
- [34]. Phatarfod R, Sudbury A. The use of a square array scheme in blood testing. *Statistics in Medicine*. 1994; 13:2337–2343. DOI: 10.1002/sim.4780132205. [PubMed: 7855467]

Table 1

Simulation results from using the ES algorithm to estimate the model in Equation (8) 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 α as described in Section 3.2. Estimated parameters and standard errors are averaged over 1000 simulated data sets. Estimated coverage probabilities are for nominal 95% Wald confidence intervals.

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_2$	$\hat{\alpha}$
0.6	500	5	Mean	-5.87	-6.95	-0.12	0.94	0.10	0.61
			SE/SD	0.92	0.96	0.92	0.92	0.92	—
			Coverage	0.93	0.96	0.94	0.94	0.94	—
	250	10	Mean	-6.09	-7.20	-0.02	1.04	0.10	0.62
			SE/SD	0.91	0.91	0.88	0.90	0.89	—
			Coverage	0.93	0.91	0.93	0.94	0.89	—
0.2	500	5	Mean	-5.99	-7.08	0.02	1.11	0.10	0.21
			SE/SD	0.99	0.98	0.98	0.91	0.95	—
			Coverage	0.95	0.95	0.95	0.94	0.95	—
	250	10	Mean	-6.01	-7.09	0.05	1.06	0.09	0.21
			SE/SD	0.84	0.87	0.84	0.83	0.84	—
			Coverage	0.91	0.92	0.95	0.94	0.93	—
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	—

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_2$	$\hat{a}$
	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 calculations for the model in Equation (8).

α	κ	I_k	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\logit(\tilde{p})$	
									$j = 1$	$j = 2$
0.6	500	5	1.344	1.802	1.150	1.218	1.482	2.403	1.272	1.388
	250	10	1.480	2.096	1.298	1.405	1.691	2.703	1.378	1.519
0.2	500	5	1.515	1.890	1.085	1.191	1.732	2.763	1.235	1.386
	250	10	1.584	1.840	1.138	1.153	1.901	2.892	1.329	1.472
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

Table 3

Parameter estimates and estimated standard errors (in parentheses) for the NPHL data described in Section 5.1. 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. We perform one joint test for each disease when evaluating race.

Disease	Term	ES algorithm		GEE	
		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 inflammatory 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 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 inflammatory 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 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 described in Section 5.2. 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 [18]. The “Overall test” column contains p-values for the Wald test $H_0 : \beta_{r1} = \beta_{r2} = \beta_{r3} = 0$ versus $H_1 : \text{at least one } \beta_{rj} \text{ not equal to } 0$, 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 $H_1 : \text{at least one } \beta_{rj} \text{ unequal}$.

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	0.303 (1.662)	<0.001	0.036
		Hepatitis B	-1.961 (1.137)		
		HIV	-4.233 (0.876)		
	Age	Syphilis	-0.099 (0.079)	0.614	0.454
		Hepatitis B	-0.020 (0.033)		
		HIV	0.004 (0.030)		
	Marital status	Syphilis	-0.667 (1.826)	0.293	0.742
		Hepatitis B	0.662 (0.540)		
		HIV	0.663 (0.505)		
	Education	Syphilis	-1.061 (0.786)	0.003	0.018
		Hepatitis B	-0.149 (0.282)		
		HIV	0.631 (0.180)		