Biometrics 000, 000–000 DOI: 000
000 0000

A mixed-effects Bayesian regression model for multivariate group testing data

Christopher S. McMahan[1]*, Chase N. Joyner[1], Joshua M. Tebbs[2], and Christopher R. Bilder[3]

[1]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

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

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

**email:* mcmaha2@clemson.edu

Summary: Laboratories use group (pooled) testing with multiplex assays to reduce the time and cost associated with screening large populations for infectious diseases. Multiplex assays test for multiple diseases simultaneously, and combining their use with group testing can lead to highly efficient screening protocols. However, these benefits come at the expense of a more complex data structure which can hinder surveillance efforts. To overcome this challenge, we develop a general Bayesian framework to estimate a mixed multivariate probit model with data arising from any group testing protocol that uses multiplex assays. In the formulation of this model, we account for the correlation between true disease statuses and heterogeneity across population subgroups, and we provide for automated variable selection through the adoption of spike and slab priors. To perform model fitting, we develop an attractive posterior sampling algorithm which is straightforward to implement. We illustrate our methodology through numerical studies and analyze chlamydia and gonorrhea group testing data collected by the State Hygienic Laboratory at the University of Iowa.

Key words: generalized linear mixed model; latent variable model; multiplex assay; multivariate probit model; pooled testing.

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

1. Introduction

The World Health Organization recently identified multiple health challenges for the next decade. These include outbreaks of novel (e.g., SARS-CoV2, etc.) and common (e.g., chlamydia, gonorrhea, etc.) diseases, a lack of access to health care, and the emergence of drug-resistant pathogens. In many instances, these challenges could be lessened with robust screening and surveillance programs that detect infected individuals and identify risk factors of disease. The primary barrier to such programs is usually the cost of implementation. One potential way to alleviate cost constraints is to amplify the use of group (pooled) testing. Group testing confers savings by testing pools of individual specimens, such as blood, urine, or swabs. Individuals in a pool that tests negatively are classified as such at the expense of a single assay, while positive pools are resolved through further testing; see Kim et al. (2007) for a review. In rare trait settings, group testing reduces costs when compared to protocols which test each specimen individually. As a result, group testing has been adopted in many areas, including infectious disease testing (Lewis et al., 2012; Krajden et al., 2014), animal health surveillance (Dhand et al., 2010), entomology (Speybroeck et al., 2012), and environmental monitoring (Heffernan et al., 2014).

Motivated by infectious disease testing practices at the State Hygienic Laboratory (SHL) at the University of Iowa, new group testing protocols using multiplex assays have been proposed recently (Tebbs et al., 2013; Hou et al., 2017; Bilder et al., 2019; Hou et al., 2020). Multiplex assays, unlike their single-disease predecessors, test for multiple diseases at once. Examples include the Procleix Ultrio Assay which tests for HIV, hepatitis B, and hepatitis C, the CDC Flu SC2 Multiplex Assay which tests for influenza A/B and SARS-CoV-2, and the Aptima Combo 2 Assay which tests for chlamydia and gonorrhea. The benefit of multiplex assays is their high-throughput potential which offers a more comprehensive assessment and a shorter turnaround time. Combining multiplex assays with group testing, it is possible

to reap the benefits of both to screen populations more efficiently. For example, the SHL tests thousands of Iowa residents each year for chlamydia and gonorrhea using group testing and the Aptima Combo 2 Assay. Annual savings are approximately $600,000, a practically significant figure for a state-run public health laboratory.

Although group testing is effective at reducing cost, its implementation can give rise to a complex data structure, especially when pools are potentially misclassified. Many authors have considered estimating a population prevalence from group testing; see Hung and Swallow (1999) for a review. More recently, the analysis of group testing data has shifted towards estimating a regression function from parametric (Vansteelandt et al., 2000; Xie, 2001), semiparametric (Wang et al., 2014; Delaigle et al., 2014), nonparametric (Delaigle and Meister, 2011; Delaigle and Hall, 2012), and Bayesian (McMahan et al., 2017; Joyner et al., 2020; Liu et al., 2021) perspectives. However, this existing work is equipped to analyze group testing data from single-disease assays. Due to the potential of coinfection and the effect of imperfect testing, extending group testing estimation methods to the multiplex setting is challenging. Initial contributions were made by Hughes-Oliver and Rosenberger (2000), Tebbs et al. (2013), and Warasi et al. (2016) to develop prevalence estimators for multiple diseases. In a regression setting, only Zhang et al. (2013) and Lin et al. (2019) have proposed approaches to model multivariate group testing data. The former considers responses from initial pools only (i.e., no retesting results are used) and the latter was designed only for the protocol in Tebbs et al. (2013). Neither approach allows for the introduction of random effects to account for heterogeneity across population subgroups.

In many large-scale screening programs, individual specimens are collected at different clinic sites throughout a geographic region and are transported to a central location for testing. Given the inherent differences among areas in a region (e.g., rural, urban, suburban, etc.) and the types of clinics providing the specimens (e.g., primary care, community health,

sexual health, etc.), it is natural to expect that heterogeneity will exist across various population subgroups. Accounting for this heterogeneity in group testing can be difficult, especially when pools are formed with individual specimens collected at different clinic sites. In fact, most existing regression methods for group testing data are not capable of accounting for this type of heterogeneity. Chen et al. (2009) and Joyner et al. (2020) have considered this issue before, but neither work is applicable in the multiplex assay setting.

In this paper, we develop a general methodology to estimate a mixed probit model (Chib and Greenberg, 1998) for multivariate group testing data. We use fixed effects to describe population-level characteristics and random effects to account for heterogeneity across population subgroups. There are several enticing features of this work. First, our methodology is completely general, allowing one to analyze data arising from any group testing protocol that uses multiplex assays. Second, our use of a multivariate model acknowledges the dependence that may exist between (or among) different diseases. Third, we cast the problem within a Bayesian framework and adopt spike and slab priors to facilitate variable selection for both the fixed and random effects. Finally, we develop a Markov chain Monte Carlo (MCMC) algorithm that consists entirely of Gibbs steps with all but one involving sampling from common distributions. Acting in unison, these features make possible the regression analysis of multiplex group testing data while accounting for its highly complex structure.

Subsequent sections are organized as follows. Section 2 provides information on the mixed multivariate probit model, modeling assumptions, deriving the observed data likelihood, and prior model elicitation. Section 3 provides an overview of the posterior sampling algorithm and data augmentation steps. Section 4 reports the results of simulation studies to assess the performance of our approach. Section 5 presents an analysis of chlamydia and gonorrhea group testing data collected by the SHL. Section 6 concludes with a summary discussion. Additional details are provided in the Supporting Information.

2. Methodology

Suppose N individuals are tested for D diseases simultaneously through a group testing protocol. We assume the protocol makes use of multiplex assays and the specimens (e.g., blood, urine, swabs, etc.) are collected from individuals at K distinct clinics. A few initial comments are in order. First, because different clinics serve different populations, a substantial amount of heterogeneity may exist across the clinic sites. Second, a group testing protocol could be performed "in-house" (i.e., at a clinic site) or at a regional laboratory like the SHL. The former would involve pooling individuals within each site, while the latter would allow for pooling individuals across the sites. Third, given the nature of most diseases tested by a multiplex assay, it is expected the true disease statuses for each individual are correlated. Our methodology accounts for all of these features among others.

Let $\widetilde{Y}_{id} = 1$ if the ith individual is truly positive for the dth disease, $\widetilde{Y}_{id} = 0$ otherwise, for $i = 1, ..., N$ and $d = 1, ..., D$. We aggregate the true disease statuses for the ith individual into the vector $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, ..., \widetilde{Y}_{iD})'$ and define $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1', ..., \widetilde{\mathbf{Y}}_N')'$. Denote by $\mathbf{x}_{id}$ and $\mathbf{t}_{id}$ the $p_d \times 1$ and $q_d \times 1$ vectors of covariates corresponding to fixed and random effects, respectively, such that $\mathbf{t}_{id}$ is a subvector of $\mathbf{x}_{id}$. We relate the individuals' true disease statuses to their covariates through a mixed multivariate probit model (Chib and Greenberg, 1998). Under this model, the distribution of $\widetilde{\mathbf{Y}}_i$ given the covariates and model parameters is

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{(i)}, \mathbf{R}) = \pi(\widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \cdots \int_{I_{iD}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega}, \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_D')'$, $\boldsymbol{\beta}_d$ is a vector of regression coefficients for the dth disease, $\boldsymbol{\gamma}_{(i)} = (\boldsymbol{\gamma}_{(i)1}', ..., \boldsymbol{\gamma}_{(i)D}')'$, $\boldsymbol{\gamma}_{(i)d}$ is a vector of random effects for ith individual associated with the dth disease, $\phi(\cdot | \boldsymbol{\eta}_i, \mathbf{R})$ is the density of a D-variate normal random vector with mean $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{iD})'$ and correlation matrix $\mathbf{R}$, $\eta_{id} = \mathbf{x}_{id}' \boldsymbol{\beta}_d + \mathbf{t}_{id}' \boldsymbol{\gamma}_{(i)d}$ is the linear predictor, and

$$I_{id} = \begin{cases} (-\infty, 0), & \text{if } \widetilde{Y}_{id} = 0 \\ [0, \infty), & \text{if } \widetilde{Y}_{id} = 1, \end{cases}$$

$d = 1, ..., D$, denote regions of integration. Note that $\mathbf{R}$ must be restricted to be a correlation matrix to ensure model identifiability; see Chib and Greenberg (1998). To account for heterogeneity across clinic sites, we adopt the convention that $\boldsymbol{\gamma}_{(i)d} = \boldsymbol{\gamma}_{kd}$ if the ith individual presents at the kth clinic. We assume the $\boldsymbol{\gamma}_{kd}$'s are iid $N(\mathbf{0}, \boldsymbol{\Sigma}_d)$ random vectors.

The model specification in (1) leads to several challenges, for example, how to identify the subset of important predictors corresponding to the random effects and specifying the covariance structure. To overcome these difficulties, we reparameterize (1) using the proposal of Chen and Dunson (2003). Using a modified Cholesky decomposition, we write the covariance matrices of the random effects as $\boldsymbol{\Sigma}_d = \boldsymbol{\Lambda}_d \mathbf{A}_d \mathbf{A}_d' \boldsymbol{\Lambda}_d$, for $d = 1, ..., D$, where $\boldsymbol{\Lambda}_d$ is a $q_d \times q_d$ diagonal matrix with nonnegative elements $\boldsymbol{\lambda}_d$ and $\mathbf{A}_d$ is a $q_d \times q_d$ lower triangular matrix with unit diagonal elements and free elements $\mathbf{a}_d = (a_{sld} \colon l = 1, ..., q_d - 1; s = l + 1, ..., q_d)'$. Aggregating $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1', ..., \boldsymbol{\lambda}_D')'$ and $\mathbf{a} = (\mathbf{a}_1', ..., \mathbf{a}_D')'$, the reparameterized model is

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \pi(\widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \cdots \int_{I_{iD}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega}, \qquad (2)$$

where $\eta_{id} = \mathbf{x}_{id}' \boldsymbol{\beta}_d + \mathbf{t}_{id}' \boldsymbol{\Lambda}_d \mathbf{A}_d \mathbf{b}_{(i)d}$ is the linear predictor under reparameterization and $\mathbf{b}_{(i)} = (\mathbf{b}_{(i)1}', ..., \mathbf{b}_{(i)D}')'$, where $\mathbf{b}_{(i)d}$ is a standardized random effect for the ith individual associated with the dth disease. To incorporate potential heterogeneity across clinic sites, we specify $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if the ith individual presents at the kth clinic and assume $\mathbf{b}_{kd} \sim N(\mathbf{0}, \mathbf{I})$.

The reparameterized model in (2) has several advantages. First, it is no longer necessary to specify or posit a prior model for the covariance matrices $\boldsymbol{\Sigma}_d$, $d = 1, ..., D$. Instead, $\boldsymbol{\Sigma}_d$ is estimated through the elements of $\boldsymbol{\Lambda}_d$ and $\mathbf{A}_d$. Second, by specifying spike and slab priors for the elements in $\boldsymbol{\lambda}_d$, we develop an automated model selection strategy that identifies predictors with associated random effects. Note that by setting a diagonal element of $\boldsymbol{\Lambda}_d = \text{diag}\{\boldsymbol{\lambda}_d\}$ equal to 0 results in the corresponding diagonal element of $\boldsymbol{\Sigma}_d$ being set to 0, which intrinsically drops the corresponding random effect from the model. Under (2), posterior inference would be relatively straightforward if the individual disease statuses $\widetilde{\mathbf{Y}}_i$

were observed (Albert and Chib, 1993; Chib and Greenberg, 1998). However, because of imperfect testing, this is not the case and the $\widetilde{\mathbf{Y}}_i$'s are best regarded as latent.

The observed data from a group testing protocol consist of diagnostic test results on a collection of pools, some of which may be of size one (i.e., individual testing). Several protocols using multiplex assays, such as those referenced in Section 1, have been proposed in the biostatistics literature. To develop a general regression methodology that accommodates all possible protocols, we track pool membership via the index set $\mathcal{P}_j$, for $j = 1, ..., J$, where $i \in \mathcal{P}_j$ if and only if the ith individual was tested in the jth pool. Therefore, the true status of the jth pool for the dth disease is $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$; i.e., the jth pool is positive for the dth disease if at least one of its members is positive for the dth disease, and these statuses are aggregated into $\widetilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, ..., \widetilde{Z}_{jD})'$. The observed test result from assaying the jth pool is $\mathbf{Z}_j = (Z_{j1}, ..., Z_{jD})'$, where $Z_{jd} = 1$ if the jth pool tests positive for the dth disease, $Z_{jd} = 0$ otherwise. To allow for imperfect testing, we let $S_{e_j:d} = P(Z_{jd} = 1 \mid \widetilde{Z}_{jd} = 1)$ and $S_{p_j:d} = P(Z_{jd} = 0 \mid \widetilde{Z}_{jd} = 0)$ denote the sensitivity and specificity, respectively, of the multiplex assay used to test the jth pool for the dth disease.

By defining $S_{e_j:d}$ and $S_{p_j:d}$ to be pool-dependent, this allows for changes in these probabilities which may be attributed to the use of different multiplex assays or other factors which could impact assay performance; e.g., the size of the jth pool, the specimen type, etc. We assume these probabilities do not vary within the strata created by cross-classifying these factors. For example, if the jth and the j'th pool are of the same size (or of a similar size), contain the same type of specimens, and are tested using the same assay, then we assume $S_{e_j:d} = S_{e_{j'}:d}$ and $S_{p_j:d} = S_{p_{j'}:d}$ for $d = 1, ..., D$. This notion is captured mathematically by defining index sets $\mathcal{I}_m$ so that $S_{e_j:d} = S_{e(m):d}$ and $S_{p_j:d} = S_{p(m):d}$ for all $j \in \mathcal{I}_m$, for $m = 1, ..., M$. We regard $S_{e(m):d}$ and $S_{p(m):d}$ as unknown which are to be estimated alongside the parameters for the fixed and random effects.

The conditional distribution of the observed testing outcomes $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_J')'$ given the covariates and the model parameters can be expressed as

$$\pi(\mathbf{Z} \mid \boldsymbol{\Theta}) = \sum_{\widetilde{\mathbf{Y}} \in \boldsymbol{\mathcal{Y}}} \left[\prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j \in \mathcal{I}_m} \left\{ S_{e(m):d}^{Z_{jd}} (1 - S_{e(m):d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p(m):d}^{1-Z_{jd}} (1 - S_{p(m):d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}} \times \prod_{i=1}^{N} \pi(\widetilde{\mathbf{Y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) \right], \quad (3)$$

where $\boldsymbol{\mathcal{Y}} = \{0, 1\}^{N \times D}$ and $\boldsymbol{\Theta}$ aggregates all model parameters. Equation (3) is derived by making mild assumptions. First, we assume testing outcomes for each disease are conditionally independent given the true pool statuses; i.e., $Z_{jd} \mid \widetilde{\mathbf{Z}}$ is independent of $Z_{j'd'} \mid \widetilde{\mathbf{Z}}$ for $(j, d) \neq (j', d')$, where $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{Z}}_1', ..., \widetilde{\mathbf{Z}}_J')'$, and the conditional distribution $\mathbf{Z} \mid \widetilde{\mathbf{Z}}$ does not depend on the covariates. Second, we assume the individual disease statuses $\widetilde{\mathbf{Y}}_i$ are conditionally independent given the covariates and the random effects. The first assumption is common in the group testing literature (e.g., see McMahan et al., 2017), while the second is ubiquitous in the literature for mixed models (e.g., see Demidenko, 2013).

Our description of the proposed model is completed by eliciting prior distributions for all parameters. To facilitate variable selection, both in the fixed and random effects components, we use spike and slab priors for $\boldsymbol{\beta}_d = (\beta_{1d}, ..., \beta_{p_d d})'$ and $\boldsymbol{\lambda}_d = (\lambda_{1d}, ..., \lambda_{q_d d})'$, for $d = 1, ..., D$. For the dth disease, prior specifications for the fixed effects are

$$\begin{aligned}
\beta_{rd} \mid v_{rd} &\sim (1 - v_{rd}) \cdot \delta_0(\beta_{rd}) + v_{rd} \cdot N(0, \phi_{rd}^2), && r = 1, ..., p_d \\
v_{rd} \mid \tau_{v_{rd}} &\sim \text{Bernoulli}(\tau_{v_{rd}}), && r = 1, ..., p_d \\
\tau_{v_{rd}} &\sim \text{beta}(a_v, b_v), && r = 1, ..., p_d,
\end{aligned}$$

whereas for the random effects,

$$\begin{aligned}
\lambda_{ld} \mid w_{ld} &\sim (1 - w_{ld}) \cdot \delta_0(\lambda_{ld}) + w_{ld} \cdot \text{TN}(0, \psi_{ld}^2, 0, \infty), && l = 1, ..., q_d \\
w_{ld} \mid \tau_{w_{ld}} &\sim \text{Bernoulli}(\tau_{w_{ld}}), && l = 1, ..., q_d \\
\tau_{w_{ld}} &\sim \text{beta}(a_w, b_w), && l = 1, ..., q_d.
\end{aligned}$$

In the prior distributions above, $\delta_0(\cdot)$ is the Dirac delta function, $\text{TN}(\mu, \sigma^2, a, b)$ denotes the truncated normal distribution that restricts a normal distribution with mean μ and variance

σ^2 to the interval (a, b), and ϕ^2_{rd}, a_v, b_v, ψ^2_{ld}, a_w, and b_w are hyperparameters. The remaining model parameters for the dth disease are the free elements $\mathbf{a}_d$ in the Cholesky decomposition matrix $\mathbf{A}_d$ and the $2M$ assay accuracy probabilities. Prior models for these parameters are

$$\begin{aligned} \mathbf{a}_d &\sim N(\mathbf{m}_d, \mathbf{C}_d) \\ S_{e(m):d} &\sim \text{beta}(a_{e(m):d}, b_{e(m):d}), \quad m = 1, ..., M \\ S_{p(m):d} &\sim \text{beta}(a_{p(m):d}, b_{p(m):d}), \quad m = 1, ..., M, \end{aligned}$$

where $\mathbf{m}_d$, $\mathbf{C}_d$, $a_{e(m):d}$, $b_{e(m):d}$, $a_{p(m):d}$, and $b_{p(m):d}$ are hyperparameters.

In the spike and slab priors, we use a Dirac delta function for the spike components, and the slab distributions are chosen to be normal and truncated normal for the fixed and random effects, respectively. Variance components of the slab distributions (ϕ^2_{rd} and ψ^2_{ld}) should be large to provide a diffuse proposal; see Wagner and Duller (2012). However, specifying the hyperparameters $\mathbf{m}_d$ and $\mathbf{C}_d$ should be done informatively (e.g., $\mathbf{m}_d = \mathbf{0}$ and $\mathbf{C}_d = 0.5\mathbf{I}$). Failing to do so results in a strong *a priori* specification for the correlation between any two random effects for the dth disease; see Chen and Dunson (2003). Finally, uninformative priors for both the mixing probability hyperparameters and the assay accuracy probabilities can be specified by setting $a_v = b_v = a_w = b_w = 1$ and $a_{e(m):d} = b_{e(m):d} = a_{p(m):d} = b_{p(m):d} = 1$, respectively. If historical information from assay validation studies is available, informative beta priors for the sensitivity and specificity parameters can be used; see Section 5.

The final parameter is the correlation matrix $\mathbf{R}$. In general, specifying a prior distribution for a correlation matrix is nontrivial due to its inherent constraints. We follow Zhang et al. (2006) and specify a joint prior for $\mathbf{R}$ and an extra variance parameter matrix $\mathbf{D}$; i.e.,

$$\pi(\mathbf{R}, \mathbf{D} \mid c_0, \mathbf{S}) \propto |\mathbf{R}|^{(c_0 - D - 1)/2} |\mathbf{D}|^{(c_0/2) - 1} \text{etr}\left(-\frac{1}{2}\mathbf{S}^{-1}\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}\right), \tag{4}$$

where $c_0 > 0$, $\mathbf{S}$ is a scale matrix, and $\text{etr}(\cdot)$ denotes the operator $\exp\{\text{tr}(\cdot)\}$. It is straightforward to show $\mathbf{W} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$ follows a Wishart distribution with c_0 degrees of freedom and scale matrix $\mathbf{S}$; i.e., $\mathbf{W} \sim \text{Wishart}(c_0, \mathbf{S})$.

3. Data Augmentation and Posterior Sampling

3.1 *Data augmentation*

Our goal is to estimate the multivariate probit model in (2) with the observed group testing responses in $\mathbf{Z}$. However, working with the observed data model $\pi(\mathbf{Z} \mid \boldsymbol{\Theta})$ in (3) is prohibitive as it involves $2^{N\times D}$ terms. We therefore propose a two-stage data augmentation strategy which leads to a convenient posterior sampling algorithm. The first stage introduces the individual disease statuses $\widetilde{Y}_{id}$ as latent random variables, producing the joint distribution

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\Theta}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\prod_{j\in\mathcal{I}_m}\left\{S_{e(m):d}^{Z_{jd}}(1-S_{e(m):d})^{1-Z_{jd}}\right\}^{\widetilde{Z}_{jd}}\left\{S_{p(m):d}^{1-Z_{jd}}(1-S_{p(m):d})^{Z_{jd}}\right\}^{1-\widetilde{Z}_{jd}} \times \prod_{i=1}^{N}\pi(\widetilde{\mathbf{Y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}).$$

The second stage introduces a latent random vector $\boldsymbol{\omega}_i = (\omega_{i1}, ..., \omega_{iD})'$ for each individual and defines $\widetilde{Y}_{id} = 1$, if $\omega_{id} \geqslant 0$, and $\widetilde{Y}_{id} = 0$ otherwise, for $d = 1, ..., D$. We regard $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_N$ to be mutually independent $N(\boldsymbol{\eta}_i, \mathbf{R})$ random vectors. This stage essentially decomposes the multivariate probit model and leads to the joint conditional distribution

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\Theta}) \propto \prod_{d=1}^{D}\prod_{m=1}^{M}\prod_{j\in\mathcal{I}_m}\left\{S_{e(m):d}^{Z_{jd}}(1-S_{e(m):d})^{1-Z_{jd}}\right\}^{\widetilde{Z}_{jd}}\left\{S_{p(m):d}^{1-Z_{jd}}(1-S_{p(m):d})^{Z_{jd}}\right\}^{1-\widetilde{Z}_{jd}} \times \prod_{i=1}^{N}|\mathbf{R}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\omega}_i-\boldsymbol{\eta}_i)'\mathbf{R}^{-1}(\boldsymbol{\omega}_i-\boldsymbol{\eta}_i)\right\}\prod_{i=1}^{N}f(\boldsymbol{\omega}_i), \quad (5)$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1', ..., \boldsymbol{\omega}_N')'$ and $f(\boldsymbol{\omega}_i) = \prod_{d=1}^{D}\{I(\omega_{id} \geqslant 0, \widetilde{Y}_{id} = 1) + I(\omega_{id} < 0, \widetilde{Y}_{id} = 0)\}$, where $I(\cdot)$ is the indicator function. Given the form of (5) and the priors elicited in Section 2, it is possible to derive closed-form full conditional distributions for the latent variables $\widetilde{Y}_{id}$ and $\boldsymbol{\omega}_i$ and all model parameters except one (the correlation matrix $\mathbf{R}$). This leads to the development of a posterior sampling algorithm that we now describe.

3.2 *Posterior sampling*

Our sampling algorithm consists entirely of Gibbs steps with all but one involving sampling from common distributions. Web Appendix A in the Supporting Information provides deriva-

tions of the following full conditional distributions and gives expressions for the parameters in these distributions. For the latent variables introduced in Section 3.1,

$$\begin{aligned}\widetilde{Y}_{id} \mid \widetilde{\mathbf{Y}}_{i(-d)}, \mathbf{Z}, \boldsymbol{\Theta} &\sim \text{Bernoulli}(p^*_{id}) \\ \boldsymbol{\omega}_i \mid \widetilde{\mathbf{Y}}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R} &\sim \text{TMN}(\boldsymbol{\eta}_i, \mathbf{R}, \mathbf{L}_i, \mathbf{U}_i),\end{aligned}$$

where $\widetilde{\mathbf{Y}}_{i(-d)}$ is the vector of all disease statuses for the ith individual excluding the dth one and TMN denotes the truncated multivariate normal distribution. The full conditional for $\widetilde{Y}_{id}$ above reiterates why our regression methodology can be used for any group testing protocol. Different protocols will produce different sets of observed group testing responses in $\mathbf{Z}$, but it suffices to keep track of the index sets $\mathcal{P}_1, ..., \mathcal{P}_J$ defined in Section 2 and the Bernoulli mean p^*_{id} does this; see Web Appendix A. For the fixed effects, the full conditional distribution of β_{rd} is degenerate at 0 if $v_{rd} = 0$, while the nonzero elements of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{\boldsymbol{v}}$, has the normal full conditional distribution

$$\boldsymbol{\beta}_{\boldsymbol{v}} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where $\mathbf{b} = (\mathbf{b}_1', ..., \mathbf{b}_K')'$, $\mathbf{b}_k = (\mathbf{b}_{k1}', ..., \mathbf{b}_{kD}')'$, $\boldsymbol{v} = (\boldsymbol{v}_1', ..., \boldsymbol{v}_D')'$, and $\boldsymbol{v}_d = (v_{1d}, ..., v_{p_d d})'$. Also,

$$\begin{aligned}v_{rd} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, \tau_{v_{rd}} &\sim \text{Bernoulli}(p_{v_{rd}}) \\ \tau_{v_{rd}} | v_{rd} &\sim \text{beta}(a_v + v_{rd}, b_v + 1 - v_{rd}),\end{aligned}$$

where $\boldsymbol{v}_{(-rd)}$ denotes the vector $\boldsymbol{v}$ with the v_{rd} entry removed. For the random effects,

$$\begin{aligned}\lambda_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \mathbf{R}, w_{ld} &\sim \text{TN}(\mu_{\lambda_{ld}} w_{ld}, \sigma^2_{\lambda_{ld}} w_{ld}, 0, \infty) \\ \mathbf{b}_k \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{R} &\sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}) \\ w_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}} &\sim \text{Bernoulli}(p_{w_{ld}}) \\ \tau_{w_{ld}} | w_{ld} &\sim \text{beta}(a_w + w_{rd}, b_w + 1 - w_{rd}),\end{aligned}$$

where $\boldsymbol{\lambda}_{(-\ell)}$ is defined in Web Appendix A. The remaining full conditionals are

$$\begin{aligned}
\mathbf{a} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b}, \mathbf{R} &\sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}) \\
S_{e(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^{\star}_{e(m):d}, b^{\star}_{e(m):d}) \\
S_{p(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^{\star}_{p(m):d}, b^{\star}_{p(m):d}).
\end{aligned}$$

To sample $\mathbf{R}$, we implement the parameter-extended Metropolis-Hastings (PX-MH) algorithm proposed by Zhang et al. (2006). This avoids having to acknowledge the inherent constraints placed on the form of $\mathbf{R}$ by sampling it jointly with $\mathbf{D}$. Moreover, the algorithm leverages the fact that $\mathbf{W} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$ is a covariance matrix to design a proposal distribution from which it is easy to sample. The PX-MH algorithm is carried out as follows.

PX-MH Algorithm

Step 1: Based on the current pair $(\mathbf{R}^{(g)}, \mathbf{D}^{(g)})$, compute $\mathbf{W}^{(g)} = {\mathbf{D}^{(g)}}^{1/2}\mathbf{R}^{(g)}{\mathbf{D}^{(g)}}^{1/2}$.

Step 2: Sample $\mathbf{W}^{\star}$ from a Wishart$(c, c^{-1}\mathbf{W}^{(g)})$ distribution.

Step 3: Compute $(\mathbf{R}^{\star}, \mathbf{D}^{\star})$ based on $\mathbf{W}^{\star} = {\mathbf{D}^{\star}}^{1/2}\mathbf{R}^{\star}{\mathbf{D}^{\star}}^{1/2}$.

Step 4: Generate $(\mathbf{R}^{(g+1)}, \mathbf{D}^{(g+1)})$ according to

$$(\mathbf{R}^{(g+1)}, \mathbf{D}^{(g+1)}) = \begin{cases} (\mathbf{R}^{\star}, \mathbf{D}^{\star}), & \text{with probability } \alpha \\ (\mathbf{R}^{(g)}, \mathbf{D}^{(g)}), & \text{otherwise.} \end{cases}$$

The acceptance probability in Step 4 is

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{R}^{\star}, \mathbf{D}^{\star} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, c_0, \mathbf{S})}{\pi(\mathbf{R}^{(g)}, \mathbf{D}^{(g)} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, c_0, \mathbf{S})} \frac{f(\mathbf{W}^{(g)} \mid \mathbf{W}^{\star})}{f(\mathbf{W}^{\star} \mid \mathbf{W}^{(g)})}\right\},$$

where $f(\cdot|\mathbf{W})$ is the proposal density based on $\mathbf{W}$ and $\pi(\mathbf{R}, \mathbf{D} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, c_0, \mathbf{S})$ is the joint posterior density of $(\mathbf{R}, \mathbf{D})$, which is proportional to

$$\pi(\mathbf{R}, \mathbf{D} \mid c_0, \mathbf{S}) \prod_{i=1}^{N} \phi(\boldsymbol{\omega}_i \mid \boldsymbol{\eta}_i, \mathbf{R}).$$

The density $f(\cdot|\mathbf{W})$ is the product of the Jacobian $\prod_{d=1}^{D} \mathbf{D}_{dd}^{(D-1)/2}$, where $\mathbf{D}_{dd}$ is the dth diagonal element of $\mathbf{D}$, and the Wishart$(c, c^{-1}\mathbf{W})$ density. The acceptance probability α is controlled by selecting c appropriately; larger values of c increase this probability.

4. Simulation Evidence

We performed various simulation experiments to examine the performance of our estimation and model selection methods. All experiments were designed to emulate the real data application in Section 5. Our primary experiment considers $N = 10000$ individuals tested for $D = 2$ diseases across $K = 50$ distinct clinic sites (200 individuals per site). For each individual, we generated the covariate vector $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})'$, where $x_{i1} \sim N(0, 1)$, $x_{i2} \sim$ Bernoulli(0.5), $x_{i3} \sim N(0, 1)$, and $x_{i4} \sim$ Bernoulli(0.5). We then set $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{t}_{i1} = \mathbf{t}_{i2} = \mathbf{x}_i^*$, where $\mathbf{x}_i^*$ denotes the vector $\mathbf{x}_i$ after being standardized. The true disease status for each individual $\widetilde{\mathbf{Y}}_i$ was generated according to

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega},$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\boldsymbol{\beta}_1 = (-2.0, -0.75, 0.5, 0, 0)'$, $\boldsymbol{\beta}_2 = (-2.5, 0, 0, 0.5, -0.25)'$, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1', \boldsymbol{\lambda}_2')'$, $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = (1, 0.75, 0.25, 0, 0)'$, $\mathbf{a} = (\mathbf{a}_1', \mathbf{a}_2')'$, where the elements of $\mathbf{a}_d$, $d = 1, 2$, are shown in Table 1, and $\mathbf{R}$ is a 2×2 correlation matrix with off diagonal elements set to 0.6. These configurations provide an overall prevalence of about 13% and 6% for diseases 1 and 2, respectively. We repeated this process independently 500 times.

For each data set, we simulated the execution of the two-stage Dorfman protocol used by the SHL and described in Tebbs et al. (2013). Under this protocol, each individual is randomly assigned to an initial pool of size four. This method of assignment allows for pools to consist of individuals from different clinics. Each pool is tested for both diseases using a multiplex assay. If a pool tests positively for either disease (or both), then each individual is retested for both diseases using the same assay. Individuals in pools that test negatively for both diseases are diagnosed as negative. The testing result for the jth pool is simulated as $Z_{jd} \mid \widetilde{Z}_{jd} \sim$ Bernoulli$\{S_{e_j:d}\widetilde{Z}_{jd} + (1 - S_{p_j:d})(1 - \widetilde{Z}_{jd})\}$, where $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$ is the true status of the jth pool. We consider two strata for the assay accuracy probabilities. The first stratum ($m = 1$) applies to initial pools, and the second stratum ($m = 2$) applies

to individuals who are retested from the first stage of testing initial pools. Based on the multiplex assay used at the SHL, we set $S_{e(1):d} = 0.95$, $S_{p(1):d} = 0.98$, $S_{e(2):d} = 0.98$, and $S_{p(2):d} = 0.99$, for $d = 1, 2$. However, when we estimated the probit model for each of the 500 group testing data sets, these quantities were treated as unknown and were assigned uniform priors; i.e., $a_{e(m):d} = b_{e(m):d} = a_{p(m):d} = b_{p(m):d} = 1$.

In the spike and slab distributions, we set $\phi^2_{rd} = \psi^2_{ld} = 100$ in the slab components to provide diffuse prior information, and we used uniform priors for all mixing weights; i.e., $a_v = b_v = a_w = b_w = 1$. The latter specification ostensibly mandates that no prior information is used in model selection for the fixed and random effects. Following Chen and Dunson (2003), we set $\mathbf{m}_d = \mathbf{0}$, $\mathbf{C}_d = 0.5\mathbf{I}$, $d = 1, 2$, to avoid specifying a strong prior correlation between any two random effects, and we set $c_0 = D+1 = 3$ and $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is a 2×2 identity matrix, to provide a diffuse prior in (4). From Section 3.2, we developed a posterior sampling algorithm to draw 100,000 MCMC iterates and retained every 10th iterate for posterior inference after discarding the first 50,000. We set the degrees of freedom in the PX-MH algorithm to be $c = 500$, which led to reasonable acceptance rates (e.g., between 20% and 40%). Standard MCMC diagnostics were performed to ensure convergence and point estimates of model parameters were determined as sample means of the posterior draws.

[Table 1 about here.]

Table 1 summarizes the results. Of primary interest are the fixed and random effects parameters β_{rd} and λ_{ld}, $r = 1, ..., 5$, $l = 1, ..., 5$, and $d = 1, 2$. For these parameters, the bias when averaged across the 500 group testing data sets is close to 0, and the sample standard deviations are small relative to the true values. Moreover, the results show our methodology can reliably identify the nonzero fixed and random effects. This can be seen from the estimated posterior probabilities of inclusion, which are unity for nearly all nonzero effects and are close to 0 when the effects are vacuous. For the remaining parameters, the

assay accuracy probabilities $S_{e(m):d}$ and $S_{p(m):d}$ are estimated nearly perfectly despite the fact that uniform priors were used, and the nuisance parameters a_{sld}, $s = 1, 2$, that is, those parameters associated with nonzero random effects, are estimated with little or no bias. Note that the inflated bias in the a_{sld} parameters, for $s = 4, 5$, is expected because these parameters are associated with null random effects; i.e., $\lambda_{41} = \lambda_{42} = \lambda_{51} = \lambda_{52} = 0$. As shown in Web Appendix A in the Supporting Information, if $\lambda_{ld} = 0$ then a_{sld} is effectively sampled from its zero-mean prior distribution for nearly all of the iterations. The correlation $\mathbf{R}_{12}$, perhaps also best regarded as a nuisance parameter, is negatively biased.

We performed three additional simulation studies that complement the findings in this section. First, we examined our estimation and model selection methods when a non-adaptive, single-stage group testing protocol was used. We observed nearly identical results to those in Table 1. Second, we compared our proposed modeling methods to the marginal (single-disease) modeling approach in Joyner et al. (2020). As expected, our multivariate approach outperforms single-disease methods in terms of estimation efficiency. Third, we performed a robustness study to examine the impact of model misspecification in the linear predictor. When a strong nonlinear relationship is present, not surprisingly, our approach can provide estimates which are biased. However, even under severe misspecification, our approach continues to reliably identify nonzero fixed and random effects. Complete details are given in Web Appendix B in the Supporting Information.

5. Iowa Data Analysis

Even at the height of the COVID-19 pandemic, the United States Centers for Disease Control and Prevention reported approximately 2.2 million new cases of chlamydia and gonorrhea in 2020 (Centers for Disease Control and Prevention, 2020), making these two of the most common sexually transmitted diseases (STDs). Both diseases are caused by bacteria, which can be passed from person to person during sexual contact. Coinfection can be common

(Creighton et al., 2003), and both bacteria are associated with the same symptoms, including painful urination and chronic pelvic pain (Workowski, 2013). However, a large percentage of infected individuals are asymptomatic which makes screening critical (Low, 2007). Both diseases can be cured with antibiotics; however, treatment is becoming more challenging as some antibiotics are now failing as a result of overuse. Given the high prevalence of both diseases, their possible long-term complications, and the looming threat of antibiotic resistance, chlamydia and gonorrhea continue to pose a serious threat to public health.

In the United States, many state-run public health laboratories have enacted screening programs which regularly test for chlamydia and gonorrhea. In Iowa, the SHL has tested thousands of residents each year dating back to the creation of the Infertility Prevention Project in 1988 (Tebbs et al., 2013). Urine and swab specimens are sent to the laboratory daily from different locations throughout the state and from different types of clinics (e.g. family planning clinics, STD clinics, etc.). Due to their higher prevalence, male specimens are tested individually, whereas most female specimens are tested by using the two-stage Dorfman protocol described in Section 4. The SHL uses the Aptima Combo 2 Assay (AC2A), which is manufactured by Hologic, Inc., to test pooled and individual specimens for both diseases simultaneously. In our analysis, we seek to identify risk factors associated with chlamydia and gonorrhea for female subjects tested at the SHL.

The data provided by our collaborators consist of testing results from female subjects in 2014. There are 4316 individual urine specimens, 416 individual cervical swab specimens, and 2286 cervical swab pool specimens (1 pool of size 2, 12 pools of size 3, and 2273 pools of size 4), as well as the additional individual test results required to resolve swab pools which test positively. These specimens represent a total of $N = 13862$ individuals from $K = 64$ clinics. In addition to the test results, several individual-level covariates were recorded, including age (in years, denoted by x_1), a race indicator ($x_2 = 1$ if Caucasian, $x_2 = 0$ otherwise), an

indicator denoting whether the subject reported a new sexual partner in the last 90 days ($x_3 = 1$ if yes), an indicator of whether the subject reported having multiple sexual partners in the last 90 days ($x_4 = 1$ if yes), an indicator of whether the subject reported sexual contact with an STD-infected partner in the previous year ($x_5 = 1$ if yes), and an indicator of whether the subject presented at a clinic with symptoms ($x_6 = 1$ if yes). We relate the individual disease statuses to these covariates through the multivariate probit model

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega},$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1', \boldsymbol{\lambda}_2')'$. In the linear predictor, we set $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{t}_{i1} = \mathbf{t}_{i2} = \mathbf{x}_i^*$, where $\mathbf{x}_i^*$ denotes the vector of covariates $\mathbf{x}_i = (1, x_{i1}, ..., x_{i6})'$ after being standardized. Standardization was used so the spike and slab distributions would have the same impact on the regression coefficients across all covariates. For each of the 64 clinics, a random effect vector $\mathbf{b}_{kd}$ is conceptualized for each disease, with the convention that $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if the ith individual was seen at the kth clinic site.

In our analysis, we used the same prior models as in Section 4 except for the assay accuracy probabilities which we model informatively. We conceptualize three strata for each disease: $S_{e(1):d}$ and $S_{p(1):d}$ for swab specimens tested individually, $S_{e(2):d}$ and $S_{p(2):d}$ for urine specimens tested individually, and $S_{e(3):d}$ and $S_{p(3):d}$ for swab specimens tested in pools. To set informative priors for these 12 parameters, we used results from AC2A validation studies, which were published in the Hologic product literature and reported in Gaydos et al. (2003). Web Appendix C in the Supporting Information reproduces these results and describes prior model construction. To estimate the model above, we used our posterior sampling algorithm to draw 100,000 MCMC iterates, retaining every 10th iterate after discarding the first half. We again used $c = 500$ as the proposal degrees of freedom in the PX-MH algorithm.

[Table 2 about here.]

[Table 3 about here.]

Tables 2 and 3 summarize the results of our analysis for chlamydia and gonorrhea, respectively, which include posterior means and standard deviations and posterior probabilities of inclusion for the fixed and random effects. The direction of the estimates of the fixed effects are consistent with known epidemiological patterns of both diseases (US Preventive Services Task Force, 2021). In particular, the risk of chlamydia tends to decrease overall with age and Caucasian females are associated with a lower risk. Having contact with a sexual partner recently diagnosed with a STD is clearly associated with increased risk for both diseases. Our analysis also identifies a random intercept parameter for both diseases and a random effect for new sexual partner associated with chlamydia, indicating evidence of heterogeneity across clinics sites. Finally, the posterior mean and standard deviation of the correlation $\mathbf{R}_{12}$ is 0.46 and 0.04, respectively. This strongly supports the use of a joint model for these data.

6. Discussion

We have formulated a Bayesian approach to model the relationship between multiple disease statuses and covariates with group testing data from multiplex assays. We estimate population-level characteristics and incorporate heterogeneity across population subgroups while identifying significant effects of each. Although we have focused on the multivariate probit model (Chib and Greenberg, 1998), other parametric approaches to model correlated binary responses may be adaptable to group testing outcomes, including logistic regression (Glonek and McCullagh, 1995). At the same time, multivariate probit models enjoy advantages such as marginal interpretation and are amendable to data augmentation strategies that lead to efficient posterior sampling. Motivated by ecological applications, Chakraborty et al. (2024) have recently investigated the multivariate probit model for high-dimensional binary responses. Future work could generalize their methods for group testing responses from multiplex assays in disease screening. For example, Koehler et al. (2018) report the development of a multiplex assay that tests for 164 different viruses, bacteria, and parasites

simultaneously. Given technological advances in modern assay development, it may soon become commonplace to test specimens for a very large number of diseases at once.

As shown in our additional simulation studies, misspecifying the linear predictor could lead to biased estimates for the fixed and random effects. To provide a more flexible approach, it might be possible to use spike and slab prior distributions to select the functional form of covariates within an additive regression modeling framework, thereby generalizing the approach in Scheipl et al. (2012) for multivariate group testing data. More broadly, a full panoply of nonparametric approaches, such as regression trees (Chipman et al., 2010) or deep learning, could be pursued to provide maximum flexibility especially if the goal is disease status prediction. Merging group testing with modern statistical learning methods is an excellent topic for future research−for single and multiple diseases.

Supporting Information

Web Appendices referenced in Sections 3-5 are available with this article at the *Biometrics* website on Wiley Online Library. We have made R programs available on our group testing research web site at `www.chrisbilder.com`.

Acknowledgements

We are grateful to the Associate Editor and two referees for their comments on an earlier version of this article. We thank Jeffrey Benfer and Kristofer Eveland at the State Hygienic Laboratory at University of Iowa. This work was funded by Grant R01 AI121351 from the National Institutes of Health and Grant OIA-1826715 from the National Science Foundation.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Bilder, C., Tebbs, J., and McMahan, C. (2019). Informative group testing for multiplex assays. *Biometrics* **75,** 278–288.

Centers for Disease Control and Prevention (2020). Sexually Transmitted Disease Surveillance 2020. Available at https://www.cdc.gov. Last accessed: May 26, 2024.

Chakraborty, A., Ou, R., and Dunson, D. (2024). Bayesian inference on high-dimensional multivariate binary responses. *Journal of the American Statistical Association* **00,** in press; doi:10.1080/01621459.2023.2260053.

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

Chen, Z. and Dunson, D. (2003). Random effects selection in linear mixed models. *Biometrics* **59,** 762–769.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85,** 347–361.

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* **4,** 266–298.

Creighton, S., Tenant-Flowers, M., Taylor, C., Miller, R., and Low, N. (2003). Co-infection with gonorrhoea and chlamydia: How much is there and what does it mean? *International Journal of STD and AIDS* **14,** 109–113.

Delaigle, A. and Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *Annals of Statistics* **40,** 131–158.

Delaigle, A., Hall, P., and Wishart, J. (2014). New approaches to non-and semi-parametric regression for univariate and multivariate group testing data. *Biometrika* **101,** 567–585.

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

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R.* John Wiley & Sons.

Dhand, N., Johnson, W., and Toribio, J. (2010). A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics* **15,** 452–473.

Gaydos, C., Quinn, T., Willis, D., Weissfeld, A., Hook, E., Martin, D., Ferrero, D., and Schachter, J. (2003). Performance of the APTIMA Combo 2 Assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in female urine and endocervical swab specimens. *Journal of Clinical Microbiology* **41,** 304–309.

Glonek, G. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B* **57,** 533–546.

Heffernan, A., Aylward, L., Toms, L., Sly, P., Macleod, M., and Mueller, J. (2014). Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology* **24,** 225–232.

Hou, P., Tebbs, J., Bilder, C., and McMahan, C. (2017). Hierarchical group testing for multiple infections. *Biometrics* **73,** 656–665.

Hou, P., Tebbs, J., Wang, D., McMahan, C., and Bilder, C. (2020). Array testing for multiplex assays. *Biostatistics* **21,** 417–431.

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

Hung, M. and Swallow, W. (1999). Robustness of group testing in the estimation of proportions. *Biometrics* **55,** 231–237.

Joyner, C., McMahan, C., Tebbs, J., and Bilder, C. (2020). From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data. *Biometrics* **76,** 913–923.

Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., and Pilcher, C. (2007). Comparison of

group testing algorithms for case identification in the presence of test error. *Biometrics* **63,** 1152–1163.

Koehler, J., Douglas, C., and Minogue, T. (2018). A highly multiplexed broad pathogen detection assay for infectious disease diagnostics. *PLoS Neglected Tropical Diseases* **12,** e0006889.

Krajden, M., Cook, D., Mak, A., Chu, K., Chahil, N., Steinberg, M., Rekart, M., and Gilbert, M. (2014). Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *Journal of Clinical Virology* **61,** 132–137.

Lewis, J., Lockary, V., and Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases* **39,** 46–48.

Lin, J., Wang, D., and Zheng, Q. (2019). Regression analysis and variable selection for two-stage multiple-infection group testing data. *Statistics in Medicine* **38,** 4519–4533.

Liu, Y., McMahan, C., Tebbs, J., Gallagher, C., and Bilder, C. (2021). Generalized additive regression for group testing data. *Biostatistics* **22,** 873–889.

Low, N. (2007). Screening programmes for chlamydial infection: When will we ever learn? *British Medical Journal* **334,** 725–728.

McMahan, C., Tebbs, J., Hanson, T., and Bilder, C. (2017). Bayesian regression for group testing data. *Biometrics* **73,** 1443–1452.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association* **107,** 1518–1532.

Speybroeck, N., Williams, C., Lafia, K., Devleesschauwer, B., and Berkvens, D. (2012). Estimating the prevalence of infections in vector populations using pools of samples.

Medical and Veterinary Entomology **26,** 361–371.

Tebbs, J., McMahan, C., and Bilder, C. (2013). Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics* **69,** 1064–1073.

US Preventive Services Task Force (2021). Screening for chlamydia and gonorrhea: US Preventive Services Task Force recommendation statement. *Journal of the American Medical Association* **326,** 949–956.

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.

Wagner, H. and Duller, C. (2012). Bayesian model selection for logistic regression models with random intercept. *Computational Statistics & Data Analysis* **56,** 1256–1274.

Wang, D., McMahan, C., Gallagher, C., and Kulasekera, K. (2014). Semiparametric group testing regression models. *Biometrika* **101,** 587–598.

Warasi, M., Tebbs, J., McMahan, C., and Bilder, C. (2016). Estimating the prevalence of multiple diseases from two-stage hierarchical pooling. *Statistics in Medicine* **35,** 3851–3864.

Workowski, K. (2013). Chlamydia and gonorrhea. *Annals of Internal Medicine* **158,** ITC2–1.

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

Zhang, B., Bilder, C., and Tebbs, J. (2013). Regression analysis for multiple-disease group testing data. *Statistics in Medicine* **32,** 4954–4966.

Zhang, X., Boscardin, J., and Belin, T. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics* **15,** 880–896.

Table 1

Simulation study. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$, the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	−0.01	0.16	1.00	$\beta_{12} = -2.5$	0.01	0.17	1.00
$\beta_{21} = -0.75$	0.01	0.15	0.99	$\beta_{22} = 0$	0.00	0.02	0.02
$\beta_{31} = 0.5$	0.00	0.05	1.00	$\beta_{32} = 0$	0.00	<0.01	0.01
$\beta_{41} = 0$	0.00	<0.01	0.01	$\beta_{42} = 0.5$	0.00	0.03	1.00
$\beta_{51} = 0$	0.00	<0.01	0.01	$\beta_{52} = -0.25$	0.00	0.03	1.00
$\lambda_{11} = 1$	0.04	0.13	1.00	$\lambda_{12} = 1$	0.05	0.14	1.00
$\lambda_{21} = 0.75$	0.02	0.09	1.00	$\lambda_{22} = 0.75$	0.02	0.09	1.00
$\lambda_{31} = 0.25$	0.00	0.05	0.99	$\lambda_{32} = 0.25$	0.00	0.05	0.99
$\lambda_{41} = 0$	0.00	<0.01	0.01	$\lambda_{42} = 0$	0.00	<0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	−0.01	0.17	–	$a_{212} = 0.5$	−0.03	0.19	–
$a_{311} = 0.2$	0.00	0.24	–	$a_{312} = 0.2$	0.00	0.25	–
$a_{321} = 0.5$	0.00	0.22	–	$a_{322} = 0.5$	−0.01	0.24	–
$a_{411} = 0.1$	−0.10	0.02	–	$a_{412} = 0.1$	−0.10	0.03	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.02	–
$a_{421} = 0.2$	−0.20	0.02	–	$a_{422} = 0.2$	−0.20	0.03	–
$a_{521} = 0.1$	−0.10	0.02	–	$a_{522} = 0.1$	−0.10	0.02	–
$a_{431} = 0.5$	−0.50	0.03	–	$a_{432} = 0.5$	−0.50	0.03	–
$a_{531} = 0.2$	−0.20	0.02	–	$a_{532} = 0.2$	−0.20	0.03	–
$a_{541} = 0.5$	−0.50	0.02	–	$a_{542} = 0.5$	−0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	0.00	0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	<0.01	–
$S_{e(2):1} = 0.98$	−0.01	0.01	–	$S_{e(2):2} = 0.98$	0.00	0.01	–
$S_{p(2):1} = 0.99$	0.00	<0.01	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–
$\mathbf{R}_{12} = 0.6$	−0.19	0.04					

Table 2

Iowa data application. Fixed and random effects results for chlamydia. The posterior mean estimate, the estimated posterior standard deviation (ESD), and the posterior probability of inclusion (PI) are shown.

Parameter	Description	Estimate	ESD	PI
β_{11}	Intercept	−1.46	0.03	1.00
β_{12}	Age	−0.23	0.02	1.00
β_{13}	Race	−0.04	0.03	0.66
β_{14}	New partner	0.02	0.03	0.29
β_{15}	Multiple partners	0.03	0.03	0.44
β_{16}	Contact with STD	0.15	0.01	1.00
β_{17}	Symptoms	0.00	0.02	0.09
λ_{11}	Intercept	0.16	0.03	1.00
λ_{12}	Age	0.00	0.01	0.01
λ_{13}	Race	0.00	<0.01	<0.01
λ_{14}	New partner	0.06	0.05	0.70
λ_{15}	Multiple partners	0.00	0.01	0.07
λ_{16}	Contact with STD	0.00	<0.01	0.01
λ_{17}	Symptoms	0.00	<0.01	<0.01
$S_{e(1):1}$	Swab individual	0.98	<0.01	–
$S_{e(2):1}$	Urine individual	0.99	<0.01	–
$S_{e(3):1}$	Swab pool	0.99	<0.01	–
$S_{p(1):1}$	Swab individual	0.98	<0.01	–
$S_{p(2):1}$	Urine individual	0.99	<0.01	–
$S_{p(3):1}$	Swab pool	0.99	<0.01	–

Table 3

Iowa data application. Fixed and random effects results for gonorrhea. The posterior mean estimate, the estimated posterior standard deviation (ESD), and the posterior probability of inclusion (PI) are shown.

Parameter	Description	Estimate	ESD	PI
β_{21}	Intercept	−2.55	0.08	1.00
β_{22}	Age	0.00	<0.01	0.01
β_{23}	Race	−0.06	0.06	0.54
β_{24}	New partner	0.00	0.01	0.01
β_{25}	Multiple partners	0.00	0.01	0.02
β_{26}	Contact with STD	0.18	0.02	1.00
β_{27}	Symptoms	0.00	0.01	0.01
λ_{21}	Intercept	0.35	0.07	1.00
λ_{22}	Age	0.01	0.02	0.07
λ_{23}	Race	0.04	0.07	0.25
λ_{24}	New partner	0.00	<0.01	<0.01
λ_{25}	Multiple partners	0.00	0.02	0.03
λ_{26}	Contact with STD	0.00	0.01	0.01
λ_{27}	Symptoms	0.00	<0.01	<0.01
$S_{e(1):2}$	Swab individual	1.00	<0.01	–
$S_{e(2):2}$	Urine individual	1.00	<0.01	–
$S_{e(3):2}$	Swab pool	1.00	<0.01	–
$S_{p(1):2}$	Swab individual	1.00	<0.01	–
$S_{p(2):2}$	Urine individual	1.00	<0.01	–
$S_{p(3):2}$	Swab pool	1.00	<0.01	–