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: Recently, laboratories have adopted group (pooled) testing protocols that make use of multiplex assays as a means to reduce the time and cost associated with screening large populations for infectious diseases. Group testing reduces cost by testing pooled specimens (e.g., blood, urine, swabs, etc.) for the presence of an infectious agent. When combined with multiplex assays, which screen for multiple diseases simultaneously, group testing offers a more timely, comprehensive, and cost effective testing protocol, when compared to traditional implementations. However, these benefits come at the expense of a far more complex data structure, which could hinder surveillance efforts. To overcome this challenge, herein we develop a general Bayesian methodology that can be used to fit a mixed multivariate probit model to data arising from any group testing protocol that makes use of a multiplex assay. In the formulation of this model, we account for the correlation between the disease statuses, the heterogeneity across population subgroups, and provide for automated variable selection through the adoption of spike and slab priors. To complete model fitting, we develop an easy to implement posterior sampling algorithm. The methodology is illustrated through a numerical study and is used to analyze chlamydia and gonorrhea screening data collected by the State Hygienic Laboratory in Iowa.

Key words: Generalized linear mixed model; Latent variable modeling; Multivariate probit model, Pooled testing; Random effects; Spike and slab prior.

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

1. Introduction

The World Health Organization has identified multiple health challenges that the world is currently facing. These range from outbreaks of both novel (e.g., SARS-CoV2) and common (e.g., HIV, chlamydia, gonorrhea) diseases, lack of access to health care, and increasing reports of drug-resistant pathogens, among others. In many instances, these challenges could be faced with robust screening and surveillance programs that are geared toward detecting infected individuals as well as identifying risk factors of the same. Regretfully, both in the United States and abroad, the primary barrier to such programs is the cost of implementation. However, a potential avenue for alleviating cost constraints could involve the adoption of group (pooled) testing. Group testing confers cost savings through testing pooled specimen formed by amalgamating specimens (e.g., blood or urine) collected from individuals. In the most basic implementations of group testing, individuals contributing to pools that test negative can be classified as such at the expense of a single diagnostic assay, while positive pools are resolved through further testing; see Kim et al. (2007) for a nice review of group testing protocols. In rare disease settings, it is easy to see that group testing can substantially reduce testing costs when compared to traditional practices which test each specimen individually. Given these potential reductions in testing cost, group testing has been adopted in a wide array of application areas; to include testing for chlamydia, gonorrhea, HIV, HBV, and HCV (Lewis et al., 2012; Kleinman et al., 2005; Sarov et al., 2007; Krajden et al., 2014), veterinary medicine (Dhand et al., 2010), entomology (Speybroeck et al., 2012), environmental monitoring (Heffernan et al., 2014), and drug discovery (Hughes-Oliver, 2006; Kainkaryam and Woolf, 2009).

Motivated by chlamydia and gonorrhea testing practices at the State Hygienic Laboratory (SHL) at the University of Iowa, several works have recently proposed new group testing protocols that make use of multiplex assays; e.g., see Tebbs et al. (2013), Hou et al. (2017), Bilder

et al. (2019), and Hou et al. (2020). Multiplex assays, unlike their traditional counterparts, test for multiple diseases simultaneously. Examples of multiplex assays include, but are not limited to, the Procleix Ultrio Assay which tests for HIV, hepatitis B, and hepatitis C, the CDC Flu SC2 multiplex assay which tests for influenza and SARS-CoV-2, and the Aptima Combo 2 Assay which tests for chlamydia and gonorrhea. The obvious benefit of multiplex assays is their high-throughput potential, which offers a more comprehensive assessment at a reduced turn around time. And while that is true, their advantages are more wide-ranging; e.g., they require a lower volume sample, which translates to a lower price-per-data point compared to traditional singleplex assays. By combing these two technologies, it would be possible to create screening programs that could screen large populations for multiple infectious diseases in a cost efficient manner. For example, the SHL currently screens thousands of Iowa residents each year for chlamydia and gonorrhea using group testing and the Aptima Combo 2 Assay. By adopting this strategy, the SHL has estimated savings to be approximately 3.1 million dollars during a recent 5-year evaluation period.

Though effective at reducing testing cost, the implementation of group testing does result in a complex data structure that is markedly more difficult to analyze. Historically, many authors have considered the problem of estimating a population prevalence based on group testing data; e.g., see Hung and Swallow (1999) for a thorough review. More recently, the paradigm has shifted to estimating regression function from a parameteric (Farrington, 1992; Vansteelandt et al., 2000; Xie, 2001), semiparametric (Wang et al., 2014; Delaigle et al., 2014), nonparametric (Delaigle and Meister, 2011; Delaigle and Hall, 2012; Wang et al., 2013), and Bayesian (McMahan et al., 2017; Joyner et al., 2020; Liu et al., 2021) perspective. However, all of the aforementioned estimation techniques are specifically designed to analyze single disease group testing data. That is, they can not accommodate data arising from a group testing protocol that makes use of a multiplex assay. Due to the high likelihood of coinfection,

especially for sexually transmitted infections, and the effects of imperfect testing, extending traditional group testing estimation techniques to the multiplex setting is a nontrivial task. In fact, only a handful of works have considered this problem. The initial contributions in this area were made by Hughes-Oliver and Rosenberger (2000), Tebbs et al. (2013), and Warasi et al. (2016) which saw the development of prevalence estimators based on multiplex group testing data. Zhang et al. (2013) and Lin et al. (2019) extended these ideas to the regression setting. However, these techniques have several key limitations. Namely, the former considers the analysis of test data taken on master pools only, while the latter was designed to analyze data arising from implementing the group testing strategy outlined in Tebbs et al. (2013). Moreover, neither of these techniques can accommodate the introduction of random effects to account for heterogeneity across population subgroups.

As a part of large screening programs, like that implemented by the SHL, individual specimen are collected at clinic sites throughout a geographic region and are transported to a central locale for testing. Given the inherent differences that exist across a region (e.g., rural, urban, etc.) and types of clinics (e.g.,primary care, community health, sexual health, etc.), it is natural to expect that heterogeneity exists across population subgroups. Accounting for this heterogeneity in group testing regression models, especially when pools are formed from individual specimen collected from different clinics, can be difficult. In fact, most previous regression methods for group testing data, such as those referenced above, are not capable of accounting for this sort of heterogeneity. To our knowledge, only two works have considered this problem (Chen and Dunson, 2003; Joyner et al., 2020), but neither are applicable in the multiplex setting.

In this paper, we develop a general methodology that can be used to fit a mixed multivariate probit model to data arising from any group testing protocol that makes use of a multiplex assay. In the formulation of this model, we make use of fixed effects to describe

the population-level characteristics and random effects to account for heterogeneity across population subgroups. We cast our problem in a Bayesian estimation framework, and adopt spike and slab priors to facilitate variable selection in both the fixed and random effects. To complete model fitting, we develop a Markov chain Monte Carlo (MCMC) sampling algorithm, that consists entirely of Gibbs steps with all but one involving sampling from common distributions. There are several novel aspects to this work. First, our modeling strategy is completely general allowing for the analysis of data arising from any group testing protocol that makes use of a multiplex assay. Second, our approach through the multivariate probit model directly acknowledges the dependence that may exist between infections. Third, through the model formulation we account for heterogeneity across population subgroups through the inclusion of random effects. Fourth, through the adoption of spike and slab priors we facilitate automated variable selection for both the fixed and random effects. These features act in unison to allow an end user to conduct the regression analysis of multiplex group testing data while directly acknowledging and accounting for its complex structure.

Subsequent sections of this article are organized as follows: Section 2 provides preliminary information regarding the proposed mixed multivariate probit model, modeling assumptions, the derivation of the observed data likelihood, and prior elicitation. Section 3 provides an overview of the posterior sampling algorithm, including data augmentation steps. Section 4 reports the results of a simulation study conducted to assess the performance of the proposed approach. Section 5 presents an analysis of chlamydia and gonorrhea testing data collected by the SHL in Iowa. Section 7 concludes with a summary discussion. Additional details required to implement the posterior sampling algorithm are provided in the Supporting Information.

2. Methodology

Suppose that N individuals are screened for D diseases simultaneously through a group testing protocol. We assume throughout that the group testing protocol makes use of a

discriminating multiplex assay and that the biospecimens (e.g., blood, urine, swabs, etc) being tested were collected from the individuals at K distinct clinics. A few comments are warranted. First, given that medical clinics serve many different functions (e.g., primary care, community health, sexual health, etc.), it is expected that a great deal of heterogeneity will exist across the clinic sites. Second, the group testing protocol could be performed in-house (i.e., at the clinical site) or at a regional lab. The former would involve pooling individuals within site, while the latter would allow for pooling across sites. Lastly, given the nature of many infectious diseases, it is expected that the disease statuses are dependent within subject. To provide a general regression framework, our proposed methodology is designed to explicitly account for all of these features among others.

To this end, let $\widetilde{Y}_{id} = 1$, for $i = 1, ..., N$ and $d = 1, ..., D$, denote the event that the ith individual is truly positive for the dth disease and $\widetilde{Y}_{id} = 0$ otherwise. For notational convenience, we aggregate the true disease statuses for the ith individual into the binary vector $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, ..., \widetilde{Y}_{iD})'$ and define $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1', ..., \widetilde{\mathbf{Y}}_N')'$. Further, let $\mathbf{x}_{id}$ and $\mathbf{t}_{id}$ denote $p_d \times 1$ 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 infection statuses to their covariates through the mixed multivariate probit model. Under this model, the conditional distribution of $\widetilde{\mathbf{Y}}_i$, given the covariates and model parameters, is

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{(i)}, \mathbf{R}) \equiv \pi(\widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \cdots \int_{I_{iD}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega}, \tag{1}$$

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 the dth disease, $\phi(\cdot|\boldsymbol{\eta}, \mathbf{R})$ is the density of a D-variate normal distribution with mean vector $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{iD})'$ and correlation matrix $\mathbf{R}$, $\eta_{id} = \mathbf{x}_{id}'\boldsymbol{\beta}_d + \mathbf{t}_{id}'\boldsymbol{\gamma}_{(i)d}$ is the usual linear predictor, and the regions of integration are

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

Note, in the model formulation $\mathbf{R}$ must be restricted to be a correlation matrix to ensure identifiability; for further discussion on this and other aspects of the multivariate probit model 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 and only if the ith individual presented at the kth clinic site, and we assume that $\boldsymbol{\gamma}_{kd} \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_d)$.

It is important to note that the current model specification leads to several challenges; namely identifying the subset of important predictors that correspond to the random effects as well as specifying the covariance structure of the same. To overcome these challenges, we reparameterize our model according to the proposal of Chen and Dunson (2003). In particular, based on a modified Cholesky decomposition, we decompose the covariance matrices of the random effects as $\boldsymbol{\Sigma}_d = \boldsymbol{\Lambda}_d \mathbf{A}_d \mathbf{A}_d' \boldsymbol{\Lambda}_d$, for $d = 1, ..., D$. Here, $\boldsymbol{\Lambda}_d$ is a $q_d \times q_d$ diagonal matrix with nonnegative diagonal elements $\boldsymbol{\lambda}_d$ and $\mathbf{A}_d$ is a $q_d \times q_d$ lower triangular matrix with unit main diagonal elements and free elements $\mathbf{a}_d = (a_{mld} \colon l = 1, ..., q_d - 1; m = l + 1, ..., q_d)'$. Aggregating $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1', ..., \boldsymbol{\lambda}_D')'$ and $\mathbf{a} = (\mathbf{a}_1', ..., \mathbf{a}_D')'$, the reparameterized model is given by

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) \equiv \pi(\widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \cdots \int_{I_{iD}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega}, \quad (2)$$

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 our reparameterization, $\mathbf{b}_{(i)d}$ is a standardized random effect for the ith individual associated with the dth disease, $\mathbf{b}_{(i)} = (\mathbf{b}_{(i)1}, ..., \mathbf{b}_{(i)D})'$, $\mathbf{b}_{(i)d} = \mathbf{b}_{kd}$ if and only if the ith individual presented at the kth clinic site, $\mathbf{b}_{kd} \sim N(\mathbf{0}, \mathbf{I})$, $\mathbf{b} = (\mathbf{b}_1', ..., \mathbf{b}_K')'$, and $\mathbf{b}_k = (\mathbf{b}_{k1}', ..., \mathbf{b}_{kD}')'$. The proposed reparameterization has several key benefits. First, it is no longer necessary to specify, or posit prior structure on, the covariance matrices $\boldsymbol{\Sigma}_d$, $d = 1, ..., D$. Instead $\boldsymbol{\Sigma}_d$ is estimated through estimating the elements of $\boldsymbol{\Lambda}_d$ and $\mathbf{A}_d$. Second, through specifying spike and slab priors for the elements of $\boldsymbol{\lambda}_d$ we can develop an automated model selection strategy that can be used to identify the subset of important predictors that correspond to the random effects. To elucidate this feature, we note that setting a diagonal element of $\boldsymbol{\Lambda}_d = \text{diag}\{\boldsymbol{\lambda}_d\}$ to zero results in the corresponding

diagonal element of $\mathbf{\Sigma}_d$ being set to zero, which effectively drops the corresponding random effect from the model. Given this model formulation, posterior estimation and inference would be relatively straight forward if the individuals' infection statuses (i.e., $\widetilde{\mathbf{Y}}_i$) were known; e.g., see Albert and Chib (1993) and Chib and Greenberg (1998). Regretfully, due to the effects of imperfect testing, this is not the case, and the individuals' statuses are best regarded as latent.

For modeling purposes, the observed data in the considered context consists of test results taken on pools as a part of a group testing protocol that uses a multiplex assay. There is a myriad of such group testing protocols that have been proposed; e.g., Tebbs et al. (2013), Hou et al. (2017), Bilder et al. (2019), and Hou et al. (2020). Moreover, many of the aforementioned protocols require individuals to be tested in multiple, possibly overlapping, pools. Thus, to develop a general regression methodology, we track pool membership via the index set $\mathcal{P}_j$, for $j = 1, ..., J$, such that $i \in \mathcal{P}_j$ if and only if the ith individual was tested in the jth pool. Using this index set, we can identify the true infection status of the jth pool for the dth disease as $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$; i.e., the jth pool is positive for the dth disease if at least one of its members is positive for the same. For the jth pool, we aggregate these statuses as $\widetilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, ..., \widetilde{Z}_{jD})'$ and define $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{Z}}_1', ..., \widetilde{\mathbf{Z}}_J')'$. Unfortunately, due to the effects of imperfect testing, the $\widetilde{\mathbf{Z}}_j$, much like the $\widetilde{\mathbf{Y}}_i$, are unobservable. Instead, we observe the test outcomes $\mathbf{Z}_j = (Z_{j1}, ..., Z_{jD})'$, where $Z_{jd} = 1$ denotes the event that the jth pool tested positive for the dth disease and $Z_{jd} = 0$ otherwise. To relate the true status of the pools to the observed test results, while accounting for imperfect testing, we assume that $S_{e_j:d} = P(Z_{jd} = 1 \mid \widetilde{Z}_{jd} = 1)$ and $S_{p_j:d} = P(Z_{jd} = 0 \mid \widetilde{Z}_{jd} = 0)$, where $S_{e_j:d}$ and $S_{p_j:d}$ are the sensitivity and specificity of the diagnostic assay used to test the jth pool.

Note, in the discussion above we allow the sensitivity and specificity of the assay to vary from pool-to-pool, thus allowing for changes in these measures that are attributable to the use

of different assays or other factors that could impact the assay's performance; e.g., specimen type, pool size (cardinality of $\mathcal{P}_j$), etc. However, these accuracy measures are not expected to vary within the testing strata defined by these factors. That is, if the jth and j'th pool were of the same size, were constructed of the same specimen type, and were tested using the same assay then we assume that $S_{e_j:d} = S_{e_{j'}:d}$ and $S_{p_j:d} = S_{p_{j'}:d}$. To capture this feature, we assume that each pool can be assigned to one of M strata and define the index sets $\mathcal{I}_m$ such 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 $m = 1, ..., M$. For our purposes, we view $S_{e(m):d}$ and $S_{p(m):d}$ as unknown quantities that have to be estimates along with the other model parameters.

Based on the hierarchy described above, and a few mild assumptions, the conditional distribution of the observed testing outcomes $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_J')'$, given the covariates and model parameters, can be expressed as

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

where $\boldsymbol{\mathcal{Y}} = \{0, 1\}^{N \times D}$ and $\mathbf{\Theta}$ aggregates the model parameters. To derive (3), we make the following assumptions. First, it is assumed that the testing outcomes for each disease are conditionally independent given the true pool statuses; i.e., $Z_{jd}|\widetilde{\mathbf{Z}}$ is independent of $Z_{j'd'}|\widetilde{\mathbf{Z}}$ for all $(j, d) \neq (j', d')$. Second, the conditional distribution $\mathbf{Z} \mid \widetilde{\mathbf{Z}}$ does not depend on the individuals' covariates. Third, the individuals' true statuses $\widetilde{\mathbf{Y}}_i$ are conditionally independent given the covariates and the random effects. Note, the first two assumptions are common among the group testing literature (e.g., see McMahan et al., 2017) while the third is ubiquitous among the mixed modeling literature (e.g., see Demidenko, 2013).

Following the tenets of the Bayesian paradigm, the proposed model is completed by specifying priors for the model parameters. To facilitate variable selection, both in the fixed and random effects components, we adopt 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 $d = 1, ..., D$. These and the other prior specifications for the dth disease are given by

$$
\begin{aligned}
\beta_{rd} \mid v_{rd} &\sim (1 - v_{rd}) \cdot \delta_0(\beta_{rd}) + v_{rd} \cdot N(0, \phi^2_{rd}), && r = 1, ..., p_d \\
v_{rd} \mid \tau_{v_{rd}} &\sim \text{Bernoulli}(\tau_{v_{rd}}), && r = 1, ..., p_d \\
\tau_{v_{rd}} &\sim \text{Beta}(a_v, b_v), && r = 1, ..., p_d \\
\lambda_{ld} \mid w_{ld} &\sim (1 - w_{ld}) \cdot \delta_0(\lambda_{ld}) + w_{ld} \cdot TN(0, \psi^2_{ld}, 0, \infty), && l = 1, ..., q_d \\
w_{ld} \mid \tau_{w_{ld}} &\sim \text{Bernoulli}(\tau_{w_{ld}}), && l = 1, ..., q_d \\
\tau_{w_{ld}} &\sim \text{Beta}(a_w, b_w), && l = 1, ..., q_d, \\
\mathbf{a}_d &\sim N(\mathbf{m}_d, \mathbf{C}_d), \\
S_{e(m):d} &\sim \text{Beta}(a_{e(m):d}, b_{e(m):d}), && m = 1, ..., M, \\
S_{p(m):d} &\sim \text{Beta}(a_{p(m):d}, b_{p(m):d}), && m = 1, ..., M,
\end{aligned}
$$

where $\mathbf{m}_d$, $\mathbf{C}_d$, a_v, a_w, b_v, b_w, ϕ^2_{rd}, ψ^2_{ld}, $a_{e(m):d}$, $b_{e(m):d}$, $a_{e(m):d}$, and $b_{e(m):d}$ are hyperparameters, $\delta_0(\cdot)$ is the Dirac delta function, and $TN(\mu, \sigma^2, a, b)$ denotes a truncated normal distribution that arises from restricting a normal distribution with mean μ and variance σ^2 to the interval (a, b). Note, in the specification of the spike and slab priors, we make use of 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, respectfully. The variance components of the slab distributions (ϕ^2_{rd}, and ψ^2_{ld}) were chosen to be large thus providing a diffuse proposal. For further details on spike and slab priors see Wagner and Duller (2012). When specifying the hyperparameters $\mathbf{m}_d$ and $\mathbf{C}_d$, care should be exercised. In particular, and counter intuitively, these hyperparameters should be specified in an informative fashion (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 within the dth disease; for further discussion see Chen and Dunson (2003). Lastly, uniformative priors for the assay accuracies can be specified by setting $a_{e(m):d} = b_{e(m):d} = a_{e(m):d} = b_{e(m):d} = 1$. Alternatively, these hyperparameters can

be chosen in a manner to leverage information from other clinical studies, as is demonstrated in the motivating data application.

Attention is now turned to eliciting a prior for $\mathbf{R}$. Recall, to ensure identifiability, $\mathbf{R}$ has to be a correlation matrix. Unlike covariance matrices, specifying priors for correlation matrices is a non-trivial task due to inherent constraints; i.e., $\mathbf{R}$ consists of bounded off-diagonal and unit main-diagonal elements. To avoid these complexities, we follow the work of Zhang et al. (2006) and specify a joint prior on $\mathbf{R}$ and an extra variance parameter matrix $\mathbf{D}$ which is given by

$$\pi(\mathbf{R}, \mathbf{D} \mid m_0, \mathbf{S}) \propto |\mathbf{R}|^{\frac{m_0-D-1}{2}} |\mathbf{D}|^{\frac{m_0}{2}-1} \text{etr}\left(-\frac{1}{2}\mathbf{S}^{-1}\mathbf{D}^{\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}\right),$$

where m_0 is the degrees of freedom, $\mathbf{S}$ is a scale matrix, and $\text{etr}(\cdot)$ denotes the operator $\exp\{\text{tr}(\cdot)\}$. Note, it is relatively straightforward to show that $\mathbf{W} = \mathbf{D}^{\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}$ obeys a Wishart distribution with degrees of freedom m_0 and scale matrix $\mathbf{S}$; i.e., $\mathbf{W} \sim \text{Wishart}(m_0, \mathbf{S})$. This realization lays the ground work for the development of a parameter-extended Metropolis-Hastings (PX-MH) algorithm that can be used to sample $\mathbf{R}$; for further discussion see Section 3.2.

3. Data Augmentation and Posterior Analysis

3.1 *Data augmentation*

It is worth noting that directly evaluating the data model outlined in (3) can be challenging due to the need to compute and sum over $2^{N\times D}$ terms. To circumvent this issue, we propose a two-stage data augmentation strategy that leads to an easy to implement posterior sampling algorithm. The first stage introduces the individuals' true statuses as latent random variables,

which leads to the following joint distribution

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\Theta}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j \in \mathcal{I}_m} \left\{ S_{e(m):d}^{Z_{jd}} (1 - S_{e(m):d})^{1-Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p(m):d}^{1-Z_{jd}} (1 - S_{p(m):d})^{Z_{jd}} \right\}^{1-\widetilde{Z}_{jd}} \times \prod_{i=1}^{N} \pi(\widetilde{\mathbf{Y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}). \tag{4}$$

In the second stage, we decompose the multivariate probit model by introducing the latent random vector $\boldsymbol{\omega}_i = (\omega_{i1}, ..., \omega_{iD})'$, for each individual, where $\boldsymbol{\omega}_i \overset{ind.}{\sim} N(\boldsymbol{\eta}_i, \mathbf{R})$ and define $\widetilde{Y}_{id} = 1$ if $\omega_{id} > 0$ and $\widetilde{Y}_{id} = 0$ otherwise. This process leads to the following joint conditional distribution

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

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1', ..., \boldsymbol{\omega}_N')'$ and $f(\boldsymbol{\omega}_i) = \prod_{d=1}^{D} I(\omega_{id} \geqslant 0, \widetilde{Y}_{id} = 1) + I(\omega_{id} < 0, \widetilde{Y}_{id} = 0)$. Given the form of (5) and the elicited priors, it is easy to ascertain the full conditionals of practically all of the model parameters. This feature leads to the easy development of a posterior sampling algorithm in the usual manner. In what follows, we provide the necessary details to construct such an algorithm.

3.2 *Posterior simulation*

Our posterior sampling algorithm consists entirely of Gibbs steps with all but one involving sampling from common distributions. Moreover, given the form of (5), the full conditionals for the latent variables introduced in Section 3.1 and all model parameters except $\mathbf{R}$ are

easily identified in the usual manner. In particular, we have the following full conditionals:

$$\widetilde{Y}_{id} \mid \widetilde{\mathbf{Y}}_{i(-d)}, \mathbf{Z}, \boldsymbol{\Theta} \sim \text{Bernoulli}(p^*_{id}),$$

$$\boldsymbol{\omega}_i \mid \widetilde{\mathbf{Y}}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R} \sim TN(\boldsymbol{\eta}_i, \mathbf{R}, \mathbf{L}_i, \mathbf{U}_i),$$

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

$$\lambda_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \mathbf{R}, w_{ld} \sim TN\{\mu_{\lambda_{ld}} w_{ld}, \sigma^2_{\lambda_{ld}} w_{ld}, 0, \infty\},$$

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

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

$$v_{rd} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, \tau_{v_{rd}} \sim \text{Bernoulli}(p_{v_{rd}}),$$

$$w_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}} \sim \text{Bernoulli}(p_{w_{ld}}),$$

$$\tau_{v_{rd}} | v_{rd} \sim \text{Beta}(a_v + v_{rd}, b_v + 1 - v_{rd}),$$

$$\tau_{w_{ld}} | w_{ld} \sim \text{Beta}(a_w + w_{rd}, b_w + 1 - w_{rd}),$$

$$S_{e(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{Beta}(a^\star_{e(m):d}, b^\star_{e(m):d}),$$

$$S_{p(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} \sim \text{Beta}(a^\star_{p(m):d}, b^\star_{p(m):d}),$$

where the specific form of the parameters of these distribution and further discussion are provided in Appendix A of the Supplementary Material. In what remains, we focus the discussion on the details required to sample $\mathbf{R}$, which is less straightforward.

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

PX-MH Algorithm

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

2. Sample $\mathbf{W}^\star$ from Wishart$(m, m^{-1}\mathbf{W}^{(g)})$.
3. Compute $(\mathbf{R}^\star, \mathbf{D}^\star)$ based on $\mathbf{W}^\star = \mathbf{D}^{\star\frac{1}{2}}\mathbf{R}^\star\mathbf{D}^{\star\frac{1}{2}}$.
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 given by

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{R}^\star, \mathbf{D}^\star \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, m_0, \mathbf{S})}{\pi(\mathbf{R}^{(g)}, \mathbf{D}^{(g)} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, m_0, \mathbf{S})} \frac{f(\mathbf{W}^{(g)} \mid \mathbf{W}^\star)}{f(\mathbf{W}^\star \mid \mathbf{W}^{(g)})}\right\},$$

where $f(\cdot|\mathbf{W})$ is the proposal density based on $\mathbf{W}$ and $\pi(\mathbf{R}, \mathbf{D} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, m_0, \mathbf{S})$ is the joint posterior density of $(\mathbf{R}, \mathbf{D})$, which is up to a constant proportional to

$$\pi(\mathbf{R}, \mathbf{D} \mid \widetilde{\mathbf{Y}}, \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, m_0, \mathbf{S}) \propto \pi(\mathbf{R}, \mathbf{D} \mid m_0, \mathbf{S}) \prod_{i=1}^{N} \phi(\boldsymbol{\omega}_i; \boldsymbol{\eta}_i, \mathbf{R}).$$

Note, the proposal density $f(\cdot|\mathbf{W})$ is the product of the Jacobian $\prod_{d=1}^{D} \mathbf{D}_{dd}^{\frac{D-1}{2}}$ and the density of a Wishart distribution with m degrees of freedom and scale matrix $m^{-1}\mathbf{W}$. Under this formulation, the acceptance rate is controlled by setting m appropriately, with larger values corresponding to an increased acceptance rate.

4. Numerical Experiments

To examine the performance of the proposed methodology, a simulation study was conducted. This study was specifically designed to mimic the salient features of the motivating data. In particular, we conceptualize a screening program tasked with testing individuals presenting at $K = 50$ clinic sites for $D = 2$ diseases. For ease of exposition we specify that 100 individuals present at each of the clinic sites, which leads to an over all sample size of $N = 5000$. This sample size is roughly a third of that available in the motivating data and therefore provides for a more than adequate benchmark for our methodology; see Section 5 for further details. For each individual, we generate a 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), and

set $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{t}_{i1} = \mathbf{t}_{i2} = \overline{\mathbf{x}}_i$, where $\overline{\mathbf{x}}_i$ denotes the vector of covariates $\mathbf{x}_i$ after being standardized. The infection status for each individual was generated according to

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

where $\boldsymbol{\beta}_1 = (-2.0, -0.75, 0.5, 0, 0)'$, $\boldsymbol{\beta}_2 = (-2.5, 0, 0, 0.5, -0.25)'$, $\boldsymbol{\lambda}_d = (1, 0.75, 0.25, 0, 0)'$, $\mathbf{a}_d = (0.5, 0.2, 0.1, 0, 0.5, 0.2, 0.1, 0.5, 0.2, 0.5)'$, and $\mathbf{R}_{12} = \mathbf{R}_{21} = 0.6$. These specifications provide for an overall prevalence rate of about 13% and 6% for disease 1 and 2, which is in keeping with the observed prevalence rate of chlamydia and gonorrhea, respectively, in our motivating data. This data generating process was used to generate 500 independent data sets.

To simulate the testing of the aforementioned data, we implement a variant of Dorfman testing that was proposed by Tebbs et al. (2013) and is currently used by the Iowa SHL. Under this protocol, each individual is randomly assigned to a group of size 4. We note that through random assignment, we are effectively pooling individuals from different clinic sites as is the case in our motivating data. Once assignment is complete, each group is tested for both diseases simultaneously. If the group tests negative for both diseases no further testing is performed and all contributing individuals are declared negative. Alternatively, if the group tests positive for either disease (or both) it is resolved through retesting the individuals one at a time for both diseases; for further discussion see Tebbs et al. (2013). Under this protocol, we simulate the test response for the jth pool as $Z_{jd} \mid \widetilde{Z}_{jd} \sim \text{Bernoulli}\{S_{e_j:d}\widetilde{Z}_{jd} + (1 - S_{p_j:d})(1 - \widetilde{Z}_{jd})\}$, where $\widetilde{Z}_{jd} = \max\{\widetilde{Y}_{id} : i \in \mathcal{P}_j\}$. To specify the sensitivity and specificity of the assay, we consider 2 testing strata. The first strata ($m = 1$) involves testing the initial pools while the second ($m = 2$) involves retesting individuals one at a time to resolve positive pools. Within these strata, we set $S_{e(1):d} = 0.95, S_{e(2):d} = 0.98, S_{p(1):d} = 0.98$, and $S_{p(2):d} = 0.99$, for $d = 1, 2$. These specifications are representative of the accuracy of the assay used by the Iowa SHL. In this study, these parameters are treated, for simplicity, as

known quantities. This should be the case when the diagnostic assay has been thoroughly validated. In contrast, when limited information about these parameters exist they can be estimated as is demonstrated in the data application; see Section 5.

We used the proposed methodology to analyze each of the 500 group testing data sets generated according to the methodology outlined above. In the implementation of our approach, we provided a diffuse specification of the slab distributions in the spike and slab priors by setting $\phi^2_{rd} = \psi^2_{ld} = 100$. Flat priors were specified for all mixing weights; i.e., $a_v = b_v = a_w = b_w = 1$. In specifying the prior for $\mathbf{a}$, we again note that this prior should be chosen to be somewhat informative to avoid specifying a strong *a priori* correlation between any two random effects; see Chen and Dunson (2003) for further discussion. Thus, we set $\mathbf{m}_0 = \mathbf{0}, \mathbf{C}_0 = 0.5\mathbf{I}$. To provide an uninformative specification of (3), we set the degrees of freedom to be $m_0 = D + 1 = 3$ and the scale matrix to be $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is a $D \times D$ identity matrix. Under these prior configurations, we used our posterior sampling algorithm to draw 100000 MCMC iterates, with every 10th iterate being retained for posterior estimation and inference after discarding the first 50000 as a burn-in. In implementing our algorithm, we set the proposal degrees of freedom in the PX-MH algorithm to be $m = 500$. This specification led to acceptable acceptance rates (e.g., between 20%-40%) in the considered setting. Standard MCMC diagnostics were conducted to insure convergence and point estimates of the model parameters were obtained as the empirical means of the posterior distributions.

Table 1 summarizes the findings from this simulations study. In particular, this table provides the the average bias and the sample standard deviation of the posterior mean estimates. Also provided are the average estimated posterior probabilities of inclusion for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Form these results, it is apparent that the proposed methodology provides both accurate point estimates and reliable inference for the fixed and random effects. That is, in

practically all cases, the empirical bias and the variability in the estimates are small relative to the true value of the corresponding parameter. Moreover, these results indicate that our methodology can reliably identify the nonzero fixed and random effects. This can be seen from the average estimated posterior probabilities of inclusion. In all cases, when the true variable is nonzero (zero) our model provides a posterior probability of inclusion that is near 1 (0). Note, it is worthwhile to point out that the inflated bias in a_{mld}, for $m = 4, 5$ and $l = 1, ..., m-1$, is expected. That is, these parameters are associated with the random effects components that are insignificant. As a part of the model fitting process, if $\lambda_{ld} = 0$ then a_{mld} is effectively sampled from its prior distribution; for further details see Appendix A of Supplementary Material. As can be seen from the average estimated posterior probabilities of inclusion associated with $\boldsymbol{\lambda}$, our model is adept at identifying the nonzero random effects. Thus, a_{mld}, for $m = 4, 5$ and $l = 1, ..., m-1$, is sampled from its prior a majority of the time accounting for the observed bias. This is in no way a limitation of the proposed methodology. In summary, this simulation study was designed to evaluate the finite sample performance of the proposed methodology in settings akin to our motivating data application. The findings of this study demonstrate the efficacy of our approach and suggests that it can be used to reliably analyze the data being collected by the Iowa SHL.

[Table 1 about here.]

5. Iowa Data Analysis

In 2019, a total of 1,808,703 and 616,392 cases of chlamydia and gonorrhea were reported to the CDC, making these diseases the most common notifiable conditions in the United States. Moreover, the incidence rates for both of these sexually transmitted infections (STIs) have steadily increased over the last decade. Both infections are caused by bacteria, which can be passed from person-to-person during sexual contact. Given that both of these bacteria

have the same modes of transmission, chlamydia and gonorrhea coinfection is common. These STIs share common symptoms, with urethritis and cervicitis being the most common among men and women, respectively. However, the large majority of infections in women are asymptomatic, and if left untreated may lead to further complications; e.g., pelvic inflammatory disease, tubal factor infertility, ectopic pregnancy, chronic pelvic pain, etc. Asymptomatic infections in men are less common, but do represent an important reservoir for transmission. Generally, these STIs are curable with antibiotics, however they are becoming more difficult to treat, with some antibiotics now failing as a result of misuse and overuse. In particular, the antibiotic resistance of these STIs has increased rapidly in recent years and has reduced treatment options. Given their prevalence, the long-term sequelae of infection, and the looming threat of antibiotic resistance, these STIs pose a serious threat to public health.

For these reasons, many states have enacted screening programs for these STIs. For example, the Iowa SHL annually screens thousands of residents for these two infections. Briefly, the SHL test specimens (e.g., urine, swab, etc.) collected from individuals at different clinics sites (e.g., family planning clinics, STD testing clinics, etc.) throughout the state. Current SHL screening protocols mandate that all male specimens and female urine specimens be tested individually while all female swab specimens are tested via a modified variant of Dorfman testing (DT); for further discussion see Tebbs et al. (2013). In all testing strata, the SHL uses the Aptima Combo 2 Assay (AC2A) to test specimens (pooled or individual) for both chlamydia and gonorrhea simultaneously.

In this analysis, we seek to identify risk factors associated with these two STIs within the female population of Iowa. The available data consist of results collected on 4316 individual urine specimens, 416 individual swab specimens, and 2286 swab master pools (1 of size 2, 12 of size 3, and 2273 of size 4), as well as the test results required to resolve the positive

master pools. Note, the group testing protocol used by the SHL mandates that all positive pools be resolved by retesting contributing individuals one-by-one for both diseases. The specimens submitted to the SHL were collected at 64 different clinic sites which were located throughout the state. In addition to the test data, several covariates were collected on each individual: age (in years, denoted by x_1), a race indicator ($x_2 = 1$ if Caucasian and $x_2 = 0$ otherwise), an indicator denoting whether the patient reported a new sexual partner in the last 90 days ($x_3 = 1$ if affirmative and $x_3 = 0$ otherwise), an indicator denoting whether the patient reported having multiple sexual partners in the last 90 days ($x_4 = 1$ if affirmative and $x_4 = 0$ otherwise), an indicator denoting whether the patient reported sexual contact with an STD-positive partner in the previous year ($x_5 = 1$ if affirmative and $x_5 = 0$ otherwise), and an indicator denoting whether the patient presented with symptoms ($x_6 = 1$ if affirmative and $x_6 = 0$ otherwise). We relate the individuals' disease statuses to the available covariate information via

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

where $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{t}_{i1} = \mathbf{t}_{i2} = \overline{\mathbf{x}}_i$, where $\overline{\mathbf{x}}_i$ denotes the vector of covariates $\mathbf{x}_i = (1, x_{i1}, ..., x_{i6})'$ after being standardized. Standardization was used so that the spike and slab distributions 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 a patient at the kth clinic site.

To complete the proposed model, all prior specifications were made in the exact same fashion as was described in Section 4 with the exception of the testing assay accuracies. In this analysis, we treat the assay accuracies as unknown quantities and conceptualize three testing 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. In total, this results in 12 sensitivity and specificity parameters that have to be

estimated. To inject knowledge about the performance of the AC2A, we specify informative priors for these parameters based on the results of the validation trials conducted by Hologic; the manufacturer of the AC2A. In particular, we specify $S_{e(m):d} \sim \text{Beta}(a_{e(m):d}, b_{e(m):d})$ and $S_{p(m):d} \sim \text{Beta}(a_{p(m):d}, b_{p(m):d})$, for $m = 1, 2, 3$ and $d = 1, 2$, where $a_{e(m):d}$, $b_{e(m):d}$, $a_{p(m):d}$, and $b_{p(m):d}$ were set, respectively, to the number of true positives, false negatives, true negatives, and false positives identified in the validation trial; see Table 2 for a summary of these values.

[Table 2 about here.]

Under these prior specifications, the proposed methodology was used to analyze these data. In the implementation, we used our posterior sampling algorithm to draw 100000 MCMC iterates, with every 10th iterate being retained for posterior estimation and inference after discarding the first 50000 as a burn-in. In implementing our algorithm, we set the proposal degrees of freedom in the PX-MH algorithm to be $m = 500$. This specification led to an acceptance rate of approximately 20%. Standard MCMC diagnostics were conducted to insure convergence and point estimates of the model parameters were obtained as the empirical means of the posterior distributions.

[Table 3 about here.]

[Table 4 about here.]

Tables 3 and 4 summarize the findings from this study for chlamydia and gonorrhea, respectively. This summary includes estimates of the posterior mean and standard deviation for all model parameters and estimates of the posterior probabilities of inclusion for the fixed and random effects. The direction of the estimates of the fixed effects are expected in light of known epidemiological patterns of chlamydia and gonorrhea infections. That is, the risk of chlamydia infection tends to decrease with age and Caucasian females are associated with a lower risk for both diseases when compared to females of other races. In

contrast, having contact with STDs is strongly associated with an increased risk. Our analysis also identifies the random intercept parameter for both diseases, and the random effect for new sexual partner associated with chlamydia, to be strongly significant indicating clear evidence of heterogeneity across the clinics throughout the state. The posterior mean and standard deviation of $\mathbf{R}_{12}$ was 0.46 and 0.04, respectively. This finding is again expected since chlamydia and gonorrhea coinfection is common. Further, this finding reinforces previous findings suggesting the same; e.g., see XXX.

6. Discussion

We have proposed a general Bayesian methodology that can be used to fit a mixed multivariate probit model to data arising from any group testing protocol that makes use of a multiplex assay. The proposed model directly acknowledges the correlation that may exist between the latent disease statuses as well as the heterogeneity that could exist across population subgroups. To facilitate automated variable selection in both the mixed and random effects, we elicit spike and slab priors. Through data augmentation steps we derive a posterior sampling algorithm that can be used to fit the proposed model. The posterior sampling algorithm consists entirely of Gibbs steps with all but one involving sampling from common distributions. The finite sample performance of the proposed approach was demonstrated via numerical experiments. Further, the proposed methodology was used to identify risk factor of chlamydia and gonorrhea through analyzing screening data collected by the State Hygienic Laboratory in Iowa. To further disseminate this work, code that implements all aspects of this work has been prepared and is available on request.

Acknowledgements

We thank Jeffrey Benfer, Dr. Lucy DesJardin, and Kristofer Eveland at the State Hygienic Laboratory (University of Iowa). This work was funded by Grant R01 AI121351 from the

National Institutes of Health. Dr. McMahan also acknowledges the support of Grant OIA-1826715 from the National Science Foundation.

References

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

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

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.

Delaigle, A. and Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *The 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.

Farrington, C. (1992). Estimating prevalence by group testing using generalized linear models. *Statistics in medicine* **11,** 1591–1597.

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. M., Bilder, C. R., and McMahan, C. S. (2017). Hierarchical group testing for multiple infections. *Biometrics* **73,** 656–665.

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

Hughes-Oliver, J. (2006). Pooling experiments for blood screening and drug discovery. *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics, Springer New York* pages 48–68.

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

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

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

Kainkaryam, R. M. and Woolf, P. J. (2009). Pooling in high-throughput drug screening. *Current opinion in drug discovery & development* **12,** 339.

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.

Kleinman, S., Strong, D., Tegtmeier, G., Holland, P., Gorlin, J., Cousins, C., Chiacchierini, R., and Pietrelli, L. (2005). Hepatitis b virus (HBV) DNA screening of blood donations in minipools with the COBAS ampliscreen HBV test. *Transfusion* **45,** 1247–1257.

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. L., Lockary, V. M., 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. S., Tebbs, J. M., Gallagher, C. M., and Bilder, C. R. (2021). Generalized additive regression for group testing data. *Biostatistics* **22,** 873–889.

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

Sarov, B., Novack, L., Beer, N., Safi, J., Soliman, H., Pliskin, J., Litvak, E., Yaari, A., and Shinar, E. (2007). Feasibility and cost–benefit of implementing pooled screening for hcvag in small blood bank settings. *Transfusion Medicine* **17,** 479–487.

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. M., McMahan, C. S., and Bilder, C. R. (2013). Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics* **69,** 1064–1073.

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.

Wang, D., Zhou, H., and Kulasekera, K. B. (2013). A semi-local likelihood regression estimator of the proportion based on group testing data. *Journal of Nonparametric Statistics* **25,** 209–221.

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

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

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

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

Table 1

Simulation results. This summary includes the average bias of the posterior mean estimates (Bias), the sample standard deviation of the estimates (SSD), and the average estimated posterior probability of inclusion (PI). The total number of individuals is $N = 5000$ with a common group size of 4.

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	0.16	1.00
$\beta_{21} = -0.75$	0.00	0.13	1.00	$\beta_{22} = 0$	0.00	0.02	0.03
$\beta_{31} = 0.5$	0.00	0.06	1.00	$\beta_{32} = 0$	0.00	0.01	0.02
$\beta_{41} = 0$	0.00	0.00	0.01	$\beta_{42} = 0.5$	0.00	0.03	1.00
$\beta_{51} = 0$	0.00	0.00	0.01	$\beta_{52} = -0.25$	0.00	0.03	1.00
$\lambda_{11} = 1$	0.04	0.14	1.00	$\lambda_{12} = 1$	0.06	0.14	1.00
$\lambda_{21} = 0.75$	0.02	0.09	1.00	$\lambda_{22} = 0.75$	0.01	0.09	1.00
$\lambda_{31} = 0.25$	-0.01	0.05	1.00	$\lambda_{32} = 0.25$	-0.01	0.05	0.99
$\lambda_{41} = 0$	0.00	0.00	0.01	$\lambda_{42} = 0$	0.00	0.00	0.01
$\lambda_{51} = 0$	0.00	0.00	0.01	$\lambda_{52} = 0$	0.00	0.01	0.01
$a_{211} = 0.5$	-0.01	0.18	–	$a_{212} = 0.5$	-0.01	0.19	–
$a_{311} = 0.2$	-0.01	0.24	–	$a_{312} = 0.2$	-0.01	0.24	–
$a_{411} = 0.1$	-0.10	0.01	–	$a_{412} = 0.1$	-0.10	0.02	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.06	–
$a_{321} = 0.5$	0.00	0.23	–	$a_{322} = 0.5$	0.00	0.21	–
$a_{421} = 0.2$	-0.20	0.01	–	$a_{422} = 0.2$	-0.20	0.01	–
$a_{521} = 0.1$	-0.10	0.01	–	$a_{522} = 0.1$	-0.10	0.02	–
$a_{431} = 0.5$	-0.50	0.01	–	$a_{432} = 0.5$	-0.50	0.01	–
$a_{531} = 0.2$	-0.20	0.02	–	$a_{532} = 0.2$	-0.20	0.01	–
$a_{541} = 0.5$	-0.50	0.01	–	$a_{542} = 0.5$	-0.50	0.01	–
$\mathbf{R}_{12} = 0.6$	-0.16	0.04					

Table 2

Data application. Presented are the true positive (TP), false negative (FN), true negative (TN), and false positive (FP) cases observed in the validation trials conducted by Hologic to examine the efficacy of the AC2A stratified by specimen type. Note, the priors for both bool and individual specimen were set using these values.

Specimen	Chlamydia				Gonorrhea			
Type	TP	FN	TN	FP	TP	FN	TN	FP
Swab	195	28	1154	12	126	17	1335	1
Urine	197	13l	1170	11	116	10l	1347	11

Table 3

Data application. This summaries the findings pertaining to chlamydia risk factors. This summary includes the posterior mean estimate (EST), posterior standard deviation estimate (ESD), and the posterior probability of inclusion (PI).

Parameter	Description	EST	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	0.00	0.00
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	0.00	0.01
λ_{17}	Symptoms	0.00	0.00	0.00
$S_{e(1):1}$	Swab individual	0.98	0.00	
$S_{e(2):1}$	Urine individual	0.99	0.00	
$S_{e(3):1}$	Swab pool	0.99	0.00	
$S_{p(1):1}$	Swab individual	0.98	0.00	
$S_{p(2):1}$	Urine individual	0.99	0.00	
$S_{p(3):1}$	Swab pool	0.99	0.00	

Table 4

Data application. This summaries the findings pertaining to gonorrhea risk factors. This summary includes the posterior mean estimate (EST), posterior standard deviation estimate (ESD), and the posterior probability of inclusion (PI).

Parameter	Description	EST	ESD	PI
β_{21}	Intercept	-2.55	0.08	1.00
β_{22}	Age	0.00	0.00	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	0.00	0.00
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	0.00	0.00
$S_{e(1):2}$	Swab individual	1.00	0.00	
$S_{e(2):2}$	Urine individual	1.00	0.00	
$S_{e(3):2}$	Swab pool	1.00	0.00	
$S_{p(1):2}$	Swab individual	1.00	0.00	
$S_{p(2):2}$	Urine individual	1.00	0.00	
$S_{p(3):2}$	Swab pool	1.00	0.00	