Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1443–1452. doi:10.1111/biom.12704.

Bayesian regression for group testing data

Christopher S. McMahan¹, Joshua M. Tebbs^{2,*}, Timothy E. Hanson², and Christopher R. Bilder³

¹Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

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

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

SUMMARY

Group testing involves pooling individual specimens (e.g., blood, urine, swabs, etc.) and testing the pools for the presence of a disease. When individual covariate information is available (e.g., age, gender, number of sexual partners, etc.), a common goal is to relate an individual's true disease status to the covariates in a regression model. Estimating this relationship is a nonstandard problem in group testing because true individual statuses are not observed and all testing responses (on pools and on individuals) are subject to misclassification arising from assay error. Previous regression methods for group testing data can be inefficient because they are restricted to using only initial pool responses and/or they make potentially unrealistic assumptions regarding the assay accuracy probabilities. To overcome these limitations, we propose a general Bayesian regression framework for modeling group testing data. The novelty of our approach is that it can be easily implemented with data from any group testing protocol. Furthermore, our approach will simultaneously estimate assay accuracy probabilities (along with the covariate effects) and can even be applied in screening situations where multiple assays are used. We apply our methods to group testing data collected in Iowa as part of statewide screening efforts for chlamydia, and we make user-friendly R code available to practitioners.

Keywords

Binary regression; Latent response; Pooled testing; Specimen pooling

1. Introduction

Group testing, also known as pooled testing, is widely used to screen individuals for sexually transmitted diseases. Originally proposed by Dorfman (1943) to detect syphilis in US soldiers during World War II, group testing involves pooling individual specimens (e.g., blood, plasma, urine, swabs, etc.) and then testing the pools for the presence of infection. Group testing confers the potential for substantial cost savings when the disease is rare,

* tebbs@stat.sc.edu.

Supplementary Materials

The Web Appendices referenced in Sections 3–7 are available with this article at the *Biometrics* website on Wiley Online Library. We have also made our R programs available on this website. A description of our programs is given in Web Appendix D.

because many individuals can be classified as disease-free when a single pool tests negatively. The infectious disease literature is replete with applications documenting the successful use of group testing, including those for HIV (Krajden et al., 2014), hepatitis B/C (Kleinman et al., 2005; Sarov et al., 2007), and chlamydia and gonorrhea (Lewis, Lockary, and Kobic, 2012). Group testing also arises naturally in other applications, including animal disease testing (Dhand, Johnson, and Toribio, 2010), environmental monitoring (Heffernan et al., 2014), and disease transmission studies involving insect vectors (Speybroeck et al., 2012).

Statistical research in group testing generally splits into two areas: estimation and case identification. This paper deals primarily with the former. Historically, the estimation problem has dealt with estimating a population-level prevalence; see Liu et al. (2012) for a review. Farrington (1992) first proposed regression analysis with group testing data using a generalized linear model with a complementary log-log link function, although his work made the restrictive assumptions that individual covariate values were identical within each pool and that testing was error free. Motivated by an HIV testing application, Vansteelandt, Goetghebeur, and Verstraeten (2000) generalized this work to incorporate imperfect tests while allowing for any link function and for any covariate composition within the pools. Modeling extensions of Vansteelandt et al. (2000) from within a parametric context include allowing for covariate measurement error (Huang, 2009), random effects (Chen, Tebbs, and Bilder, 2009), and external biomarker information (McMahan, Tebbs, and Bilder, 2013). Delaigle and Meister (2011), Delaigle, Hall, and Wishart (2014), and Delaigle and Hall (2015) have recently championed nonparametric approaches.

All of the regression methods described so far in this article share one common characteristic, namely, each uses only the initial (master) pool responses for estimation. One might regard this to be a limitation because in most disease testing applications, positive master pools are resolved in stages to distinguish positive individuals from negative ones (i.e., the case identification problem) and, in some cases, individuals initially declared negative are retested to isolate false-negatives (Gastwirth and Johnson, 1994; Krajden et al., 2014). If additional retesting responses are available, including them in the analysis can sharpen one's inference for the covariate effects. However, by comparison, far fewer regression methods are available that can incorporate this extra information. The most notable contribution is Xie (2001), who developed an expectation-maximization (EM) algorithm by viewing the true individual statuses as missing data. Zhang, Bilder, and Tebbs (2013) implemented Xie's (2001) EM algorithm for specific case identification protocols, showing that inference from group testing can be more efficient than individual testing despite requiring fewer tests.

In this article, we provide a general Bayesian framework to estimate regression models with group testing data. Our regression methods enjoy three advantages when compared to previous work in this area. First, taking a Bayesian approach is natural given the sequential manner in which screening data accumulate over time. Our approach allows one to incorporate historical information about disease prevalence whereas existing methods do not. Second, previous work in group testing regression requires assay accuracy probabilities to be known in advance. In contrast, by incorporating assay accuracy information from

preliminary studies, we allow for simultaneous estimation of the regression coefficients and accuracy probabilities, even when multiple assays are used. Finally, our framework can seamlessly incorporate all of the data collected from *any* type of group testing protocol. In other words, regardless of how the testing results arise (i.e., from master pools, from subsets of master pools, or from individuals) and regardless of the individual covariate composition within the pools, there are no alterations to our underlying estimation procedure. From our survey of the group testing literature, we can find no other regression approach that is this flexible.

Subsequent sections of this article are organized as follows. In Section 2, we introduce notation and describe modeling assumptions. In Section 3, we discuss the use of data augmentation that facilitates our Bayesian framework and describe posterior sampling when assay accuracy probabilities are known. In Section 4, we consider the case where accuracy probabilities are unknown and are estimated simultaneously along with the regression parameters. In Section 5, we use simulation to investigate the quality of our regression methods under a variety of group testing protocols and prior models. In Section 6, we illustrate our techniques using chlamydia data collected as part of statewide screening efforts in Iowa. In Section 7, we conclude with a discussion and describe future research.

2. Preliminaries

Suppose group testing is used to screen N individuals for a binary characteristic, such as disease status. Let $\tilde{Y}_i=1$ if the i th individual is truly positive, $\tilde{Y}_i=0$ otherwise, $i=1, 2, \dots, N$. In some group testing protocols, such as master pool testing (Vansteelandt et al., 2000), individual testing is never performed. In other protocols, such as those designed for case identification (Kim et al., 2007), some subjects are retested in subpools and/or individually to complete the classification process. Because imperfect assays prevent one from observing the true individual statuses in either case, the $\tilde{Y}_i$'s are best regarded as latent.

To keep our methodology as general as possible; i.e., to incorporate data from any group testing protocol, define $\mathcal{P}_j \subset \{1, 2, \dots, N\}$ to be the set of indices identifying which individuals belong to the j th pool, for $j=1, 2, \dots, J$. The $\mathcal{P}_j$ notation keeps track of pool membership, noting that the same individual could reside in multiple pools throughout the testing process. Our only requirement is that $\cup_j \mathcal{P}_j = \{1, 2, \dots, N\}$. Note that $\mathcal{P}_j$ contains a single index if there is only one individual in the pool; i.e., standard diagnostic screening.

Denote by $\tilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \tilde{Y}_i > 0)$ the true (latent) status of the j th pool; i.e., $\tilde{Z}_j=1$ if the j th pool contains at least one truly positive individual, $\tilde{Z}_j=0$ otherwise. Similarly, let Z_j denote the observed testing response for the j th pool. Allowing for imperfect tests, let

$S_{e_j} = \text{pr}(Z_j=1 | \tilde{Z}_j=1)$ and $S_{p_j} = \text{pr}(Z_j=0 | \tilde{Z}_j=0)$ denote the sensitivity and specificity, respectively, for the j th pool. We allow S_{e_j} and S_{p_j} to be pool-specific, reflecting that different pools could have different accuracies depending on how large the pools are and what type of assay is used.

We assume the i th individual's true status is related to the linear predictor $\mathbf{x}_i'\beta$ through a monotone differentiable link function $g(\cdot)$; i.e., $\text{pr}(\tilde{Y}_i=1|\mathbf{x}_i, \beta)=g^{-1}(\mathbf{x}_i'\beta)$, where $\mathbf{x}_i=(1, x_{i1}, \dots, x_{ir})'$ is an $(r+1) \times 1$ vector of covariates observed for the i th individual and $\beta=(\beta_0, \beta_1, \dots, \beta_r)'$ is a vector of regression coefficients. Let $\mathbf{Z}=(Z_1, Z_2, \dots, Z_J)'$ denote the vector of observed testing responses. Assuming the $\tilde{Y}_i$'s are conditionally independent given the covariates, the conditional distribution of $\mathbf{Z}$ for any group testing protocol can be written as

$$\pi(\mathbf{Z}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta) = \sum_{\tilde{\mathbf{Y}} \in \{0,1\}^N} \left\{ \prod_{j=1}^J \{S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j}\}^{\tilde{Z}_j} \{(1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j}\}^{1-\tilde{Z}_j} \right. \\ \left. \times \prod_{i=1}^N \{g^{-1}(\mathbf{x}_i'\beta)\}^{\tilde{Y}_i} \{1 - g^{-1}(\mathbf{x}_i'\beta)\}^{1-\tilde{Y}_i} \right\}, \quad (1)$$

where $\mathbf{S}_e=(S_{e1}, S_{e2}, \dots, S_{e_J})'$, $\mathbf{S}_p=(S_{p1}, S_{p2}, \dots, S_{p_J})'$, $\mathbf{X}=(\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_N)'$ is an $N \times (r+1)$ matrix that aggregates all of the individual covariates, and $\tilde{\mathbf{Y}}=(\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_N)'$. In writing Equation (1), we assume the testing responses in $\mathbf{Z}$ are conditionally independent given $\tilde{\mathbf{Y}}$ and that the conditional distribution $\mathbf{Z}|\tilde{\mathbf{Y}}$ does not depend on the covariates. These assumptions are common in the group testing literature when allowing for imperfect tests; see Sections 5 and 7 for further discussion.

Because the sum is taken over all 2^N possible true statuses of $\tilde{\mathbf{Y}}$, one can see the potential difficulty in evaluating Equation (1). When testing only the master pools for estimation purposes, $\pi(\mathbf{Z}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta)$ collapses substantially making direct evaluation feasible. However, when subsequent retesting results on subpools or individuals are introduced, $\pi(\mathbf{Z}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta)$ can quickly become intractable; furthermore, its form changes depending on which group testing protocol is used (Xie, 2001). Unfortunately, this has limited the usefulness of applying group testing regression methods in practice—either the methods use limited information (master pool results only) or they are inextricably linked to specific protocols that may not be used in a given application. Our Bayesian approach avoids these computational challenges and incorporates any type of group testing data.

3. Estimation

In this section, we assume the testing accuracies S_{ej} and S_{pj} are known and thus estimation of β is of central interest. To provide a general framework, we introduce the individuals' true statuses $\tilde{Y}_i$ as latent random variables so that the conditional distribution of $\mathbf{Z}$ and $\tilde{\mathbf{Y}}$ is

$$\pi(\mathbf{Z}, \tilde{\mathbf{Y}}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta) = \prod_{j=1}^J \{S_{ej}^{Z_j} (1 - S_{ej})^{1-Z_j}\}^{\tilde{Z}_j} \{(1 - S_{pj})^{Z_j} S_{pj}^{1-Z_j}\}^{1-\tilde{Z}_j} \\ \times \prod_{i=1}^N \{g^{-1}(\mathbf{x}_i'\beta)\}^{\tilde{Y}_i} \{1 - g^{-1}(\mathbf{x}_i'\beta)\}^{1-\tilde{Y}_i}. \quad (2)$$

Let $\pi(\boldsymbol{\beta})$ denote the prior distribution for $\boldsymbol{\beta}$ so that the joint posterior distribution of $\tilde{\mathbf{Y}}$ and $\boldsymbol{\beta}$ is $\pi(\tilde{\mathbf{Y}}, \boldsymbol{\beta} | \mathbf{Z}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}) \propto \pi(\mathbf{Z}, \tilde{\mathbf{Y}} | \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta})$. Unless otherwise stated, we take $\pi(\boldsymbol{\beta})$ to be a multivariate normal density with mean $\mathbf{a}$ and covariance matrix $\mathbf{R}$ so that

$$\pi(\tilde{\mathbf{Y}}, \boldsymbol{\beta} | \mathbf{Z}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}) \propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1}(\boldsymbol{\beta} - \mathbf{a}) + \sum_{i=1}^N \tilde{\mathbf{Y}}_i \theta_i - \mathbf{b}(\theta_i) \right\} \\ \times \prod_{j=1}^J \{S_{e_j}^{Z_j} (1 - S_{e_j})^{1-Z_j}\}^{\tilde{Z}_j} \{(1 - S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{1-\tilde{Z}_j},$$

where $\theta_i = \log[g^{-1}(\mathbf{x}'_i \boldsymbol{\beta}) / \{1 - g^{-1}(\mathbf{x}'_i \boldsymbol{\beta})\}]$ and $\mathbf{b}(x) = \log\{1 + \exp(x)\}$. If the latent statuses $\tilde{\mathbf{Y}}$ were observed, sampling $\boldsymbol{\beta}$ could utilize any Bayesian method for binary regression (see, e.g., Albert and Chib, 1993; Basu and Mukhopadhyay, 2000). However, many of these sampling techniques require an additional layer of data latency, which could be computationally expensive when N is large. We use Gamerman's (1997) Metropolis-Hastings algorithm which was designed for generalized linear models and is computationally simpler. Under our prior specification, the full conditional posterior density of $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta} | \tilde{\mathbf{Y}}, \mathbf{X}) \propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1}(\boldsymbol{\beta} - \mathbf{a}) + \sum_{i=1}^N \tilde{\mathbf{Y}}_i \theta_i - \mathbf{b}(\theta_i) \right\}. \quad (3)$$

To sample from $\pi(\boldsymbol{\beta} | \tilde{\mathbf{Y}}, \mathbf{X})$, we consider a proposal density that is computed based on the previously sampled value of $\boldsymbol{\beta}$. Following McCullagh and Nelder (1989), define the vector of latent observations $\tilde{\mathbf{Y}}(\boldsymbol{\beta})$ and the diagonal matrix of weights $\mathbf{W}(\boldsymbol{\beta})$ to have individual components $\tilde{Y}_i(\boldsymbol{\beta}) = \mathbf{x}'_i \boldsymbol{\beta} + \{\tilde{Y}_i - g^{-1}(\mathbf{x}'_i \boldsymbol{\beta})\} g' (g^{-1}(\mathbf{x}'_i \boldsymbol{\beta}))$ and $W_i^{-1}(\boldsymbol{\beta}) = b''(\theta_i) \{g' (g^{-1}(\mathbf{x}'_i \boldsymbol{\beta}))\}^2$, respectively, for $i = 1, 2, \dots, N$. The proposal density based on the previously sampled value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(t-1)}$, is multivariate normal with mean $\mathbf{m}^{(t)} = \{\mathbf{R}^{-1} + \mathbf{X}' \mathbf{W}(\boldsymbol{\beta}^{(t-1)}) \mathbf{X}\}^{-1} \{\mathbf{R}^{-1} \mathbf{a} + \mathbf{X}' \mathbf{W}(\boldsymbol{\beta}^{(t-1)}) \tilde{\mathbf{Y}}(\boldsymbol{\beta}^{(t-1)})\}$ and covariance matrix $\mathbf{C}^{(t)} = \{\mathbf{R}^{-1} + \mathbf{X}' \mathbf{W}(\boldsymbol{\beta}^{(t-1)}) \mathbf{X}\}^{-1}$.

3.1. Posterior Sampling

The full conditional posterior density of $\boldsymbol{\beta}$ in Equation (3) depends on the $\tilde{\mathbf{Y}}_i$'s, which are not observed. However, the observed group testing responses $\mathbf{Z}$ and the individual covariates certainly provide us with information about the $\tilde{\mathbf{Y}}_i$'s. Using the conditional distribution in Equation (2), we derive the full conditional posterior of $\tilde{\mathbf{Y}}_i$ to be $\tilde{\mathbf{Y}}_i | \mathbf{Z}, \tilde{\mathbf{Y}}_{-i}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}$, $\boldsymbol{\beta} \sim \text{Bernoulli}\{p_{i1}^* / (p_{i0}^* + p_{i1}^*)\}$, where $\tilde{\mathbf{Y}}_{-i} = (\tilde{Y}_1, \dots, \tilde{Y}_{i-1}, \tilde{Y}_{i+1}, \dots, \tilde{Y}_N)'$,

$$p_{i1}^* = g^{-1}(\mathbf{x}_i' \beta) \prod_{j \in \mathcal{A}_i} S_{e_j}^{Z_j} (1 - S_{e_j})^{1-Z_j}$$

$$p_{i0}^* = \{1 - g^{-1}(\mathbf{x}_i' \beta)\} \prod_{j \in \mathcal{A}_i} \{S_{e_j}^{Z_j} (1 - S_{e_j})^{1-Z_j}\}^{I(\sum_{i' \in \mathcal{P}_{ij}} \tilde{Y}_{i'} > 0)} \{(1 - S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{I(\sum_{i' \in \mathcal{P}_{ij}} \tilde{Y}_{i'} > 0)},$$

$\mathcal{A}_i = \{j: i \in \mathcal{P}_j\}$, and $\mathcal{P}_{ij} = \{i' \in \mathcal{P}_j: i' \neq i\}$. Examining this conditional distribution closely reveals why our Bayesian framework can handle data from any group testing protocol; i.e., all one needs to do is simply keep track of which individuals are in which pools. Furthermore, this distribution suggests the natural use of a Gibbs sampler to “recover” the latent $\tilde{Y}_i$ ’s in any application. A description of our posterior sampling algorithm is now given.

POSTERIOR SAMPLING ALGORITHM

-
- 1 Initialize $\beta^{(0)}$ and $\tilde{Y}_i^{(0)} = 0, i = 1, 2, \dots, N$. Set $t = 1$.
 - 2 For $i = 1, 2, \dots, N$, sample $\tilde{Y}_i^{(t)} \sim \text{Bernoulli}\{p_{i1}^* / (p_{i0}^* + p_{i1}^*)\}$, where p_{i0}^* and p_{i1}^* are evaluated at $\tilde{\mathbf{Y}}^{(t)} = (\tilde{Y}_1^{(t)}, \dots, \tilde{Y}_{i-1}^{(t)}, \tilde{Y}_{i+1}^{(t-1)}, \dots, \tilde{Y}_N^{(t-1)})'$ and $\beta^{(t-1)}$. Set $\tilde{\mathbf{Y}}^{(t)} = (\tilde{Y}_1^{(t)}, \tilde{Y}_2^{(t)}, \dots, \tilde{Y}_N^{(t)})'$.
 - 3 Sample $\beta^{(t)}$ from $\pi(\beta | \tilde{\mathbf{Y}}^{(t)}, \mathbf{X})$ according to the following steps:
 - a. Sample β^* from the proposal density $\mathcal{N}_{r+1}(\mathbf{m}^{(t-1)}, \mathbf{C}^{(t-1)})$
 - b. Accept β^* as $\beta^{(t)}$ with probability $\alpha(\beta^{(t-1)} | \beta^*)$
 - c. If β^* is not accepted, set $\beta^{(t)} = \beta^{(t-1)}$.
 - 4 Set $t = t + 1$ and return to Step 2.
-

The acceptance probability in Step 3b is

$$\alpha(\beta^{(t-1)} | \beta^*) = \min \left\{ 1, \frac{\pi(\beta^* | \tilde{\mathbf{Y}}^{(t)}, \mathbf{X}) \mathbf{q}(\beta^{(t-1)} | \beta^*)}{\pi(\beta^{(t-1)} | \tilde{\mathbf{Y}}^{(t)}, \mathbf{X}) \mathbf{q}(\beta^* | \beta^{(t-1)})} \right\},$$

where $\mathbf{q}(\beta^* | \beta^{(t-1)})$ is the density in Step 3a evaluated at β^* . Likewise, $\mathbf{q}(\beta^{(t-1)} | \beta^*)$ is the $\mathcal{N}_{r+1}(\mathbf{m}^*, \mathbf{C}^*)$ density evaluated at $\beta^{(t-1)}$, where $\mathbf{m}^*$ and $\mathbf{C}^*$ depend on β^* . Regardless of which group testing protocol is used, this posterior sampling algorithm is extremely fast because all unknown quantities are updated using standard distributions. We illustrate this in Section 5 using three well-known protocols for estimation and case identification.

3.2. Prior Elicitation for β

We describe three approaches to select a prior for the regression parameter β . The first is motivated by the power prior class espoused by Ibrahim and Chen (2000). This class

assumes the availability of “historical data” which is weighted according to how much influence the analyst chooses them to have. In this spirit, we specify $\pi(\boldsymbol{\beta})$ to be a multivariate normal density with mean $\mathbf{a}$ and covariance matrix $\mathbf{R}$, where $\mathbf{R} = \mathbf{a}_0^{-1} \mathbf{R}_0$. The quantities $\mathbf{a}$ and $\mathbf{R}_0$ are determined from the historical data and $a_0 \in [0, 1]$ controls the amount of influence these data have on the current analysis. Note that as $a_0 \rightarrow 0$, the historical data are weighted less and the $a_0 = 0$ case corresponds to specifying $\pi(\boldsymbol{\beta}) \propto 1$, an improper prior. Taking $a_0 = 1$ weights the historical and current data equally. This class of priors is sensible for large-scale disease testing applications, because testing outcomes from previous periods of screening can be treated as historical data.

In contrast, one could specify a prior distribution for the individual disease probabilities p_i which, in turn, induces a prior for $\boldsymbol{\beta}$. The population-averaged approach of Hanson, Branscum, and Johnson (2014) constructs a normal prior for $\boldsymbol{\beta}$ by using the distribution of the probabilities $p_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$ while assuming the covariate $\mathbf{x}_i$ is independent of $\boldsymbol{\beta}$. In particular, suppose the p_i 's are regarded as independent and identically distributed from a beta distribution with parameters a_p and b_p . For logistic regression, Hanson et al. (2014) show that this gives rise to the g-prior $\boldsymbol{\beta} \sim \mathcal{N}_{r+1}(\mathbf{a}, \mathbf{R})$, where $\mathbf{a} = \{\delta(a_p) - \delta(b_p)\} \mathbf{e}_1$ and $\mathbf{R} = \mathcal{N}[\delta'(a_p) + \delta'(b_p)](\mathbf{X}'\mathbf{X})^{-1}$. In these expressions, $\mathbf{e}_1$ is the standard unit vector whose first element is 1 and $\delta(\cdot)$ and $\delta'(\cdot)$ are the digamma and trigamma functions, respectively.

Finally, the approach taken by Bedrick, Christensen, and Johnson (1996) induces a prior for $\boldsymbol{\beta}$ by first breaking up the covariate space into r sub-populations. Let $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_r$ denote linearly independent covariate vectors that are “representative” of the sub-populations, and let $\tilde{p}_k = g^{-1}(\tilde{\mathbf{x}}_k' \boldsymbol{\beta})$, $k = 1, 2, \dots, r$, denote the disease probability in the k th sub-population. Bedrick et al. (1996) show that if $\tilde{p}_k \sim \text{beta}(a_{\tilde{p}_k}, b_{\tilde{p}_k})$ independently, the induced prior is $\pi(\boldsymbol{\beta}) \propto \prod_{k=1}^r \{g^{-1}(\tilde{\mathbf{x}}_k' \boldsymbol{\beta})\}^{a_{\tilde{p}_k}} \{1 - g^{-1}(\tilde{\mathbf{x}}_k' \boldsymbol{\beta})\}^{b_{\tilde{p}_k}}$ under logistic regression, with suitable adjustments made for other link functions. Using this prior requires a slight modification to our posterior sampling algorithm in Section 3.1; see Web Appendix A.

4. Unknown Assay Accuracies

We now extend our regression techniques to allow S_{e_j} and S_{p_j} to be unknown. Existing group testing regression methods largely assume these probabilities are known and most methods further assume that S_{e_j} and S_{p_j} do not depend on the size of the pool; i.e., $S_{e_j} = S_e$ and $S_{p_j} = S_p$, for $j = 1, 2, \dots, J$. In practice, S_e and S_p are rarely ever known exactly; instead, they are usually estimated on the basis of pilot studies performed by assay manufacturers and published in the product literature. Therefore, not only are these values subject to natural sampling variability, but such studies may use specimens that are not representative of those in the current screening population. Furthermore, published estimates of S_e and S_p are usually for individual testing outcomes—not for results based on testing pools.

Allowing the assay accuracy probabilities to be unknown creates very few additional challenges from within our Bayesian framework—even when multiple assays are used. Reparameterizing, let $S_{\alpha(l)}$ and $S_{\beta(l)}$ denote the sensitivity and specificity associated with the

l th assay, $l = 1, 2, \dots, L$, and define $\mathcal{M}(l) = \{j: \text{the } l\text{th assay was used to test the } j\text{th pool}\}$. This notation allows for different assays to be used throughout the testing process; e.g., screening and confirmatory tests. This notation also allows for the same assay to be used on pools of different sizes where $S_{e(l)}$ and $S_{p(l)}$ change depending on the size of the pool. Under this reparameterization, the conditional distribution of $\mathbf{Z}$ and $\tilde{\mathbf{Y}}$ can be written as

$$\pi(\mathbf{Z}, \tilde{\mathbf{Y}} | \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta) = \prod_{l=1}^L \prod_{j \in \mathcal{M}(l)} \{ (S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j})^{\tilde{Z}_j} \{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \}^{1-\tilde{Z}_j} \} \\ \times \prod_{i=1}^N \{ g^{-1}(\mathbf{x}'_i \beta) \}^{\tilde{Y}_i} \{ 1 - g^{-1}(\mathbf{x}'_i \beta) \}^{1-\tilde{Y}_i}, \quad (4)$$

where $\mathbf{S}_e = (S_{e(1)}, S_{e(2)}, \dots, S_{e(L)})'$ and $\mathbf{S}_p = (S_{p(1)}, S_{p(2)}, \dots, S_{p(L)})'$. To acknowledge uncertainty in the assay accuracies, we introduce independent prior distributions $\pi(S_{e(l)})$ and $\pi(S_{p(l)})$, $l = 1, 2, \dots, L$, so that the posterior distribution is

$$\pi(\tilde{\mathbf{Y}}, \mathbf{S}_e, \mathbf{S}_p, \beta | \mathbf{Z}, \mathbf{X}) \propto \pi(\mathbf{Z}, \tilde{\mathbf{Y}} | \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta) \pi(\beta) \prod_{l=1}^L \pi(S_{e(l)}) \pi(S_{p(l)}).$$

The form of the conditional distribution in Equation (4) suggests that $\text{beta}(a_{S_{e(l)}}, b_{S_{e(l)}})$ and $\text{beta}(a_{S_{p(l)}}, b_{S_{p(l)}})$ priors are natural choices for $\pi(S_{e(l)})$ and $\pi(S_{p(l)})$, respectively. Under this assumption, the full conditional posterior distributions are also beta, specifically,

$$S_{e(l)} | \mathbf{Z}, \tilde{\mathbf{Y}} \sim \text{beta}(\mathbf{a}_{S_{e(l)}}^*, \mathbf{b}_{S_{e(l)}}^*) \text{ and } S_{p(l)} | \mathbf{Z}, \tilde{\mathbf{Y}} \sim \text{beta}(\mathbf{a}_{S_{p(l)}}^*, \mathbf{b}_{S_{p(l)}}^*), \quad l = 1, 2, \dots, L,$$

where $a_{S_{e(l)}}^* = a_{S_{e(l)}} + \sum_{j \in \mathcal{M}(l)} Z_j \tilde{Z}_j$, $b_{S_{e(l)}}^* = b_{S_{e(l)}} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j) \tilde{Z}_j$, $a_{S_{p(l)}}^* = a_{S_{p(l)}} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j) (1 - \tilde{Z}_j)$, and $b_{S_{p(l)}}^* = b_{S_{p(l)}} + \sum_{j \in \mathcal{M}(l)} Z_j (1 - \tilde{Z}_j)$. The full conditional posterior distributions of β and the individual true statuses $\tilde{Y}_i$ are the same as those in Section 3.1. The expressions for p_{i1}^* and p_{i0}^* in the distribution of $\tilde{Y}_i | \mathbf{Z}, \tilde{\mathbf{Y}}_{-i}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta$ change in appearance to reflect the assay accuracy reparameterization we made in this section.

We provide a complete description of the posterior sampling algorithm, now for $\beta, \mathbf{S}_e$, and $\mathbf{S}_p$ in Web Appendix B. The algorithm is identical to the one given in Section 3.1 except that an extra Gibbs sampling step is added for the testing accuracies.

5. Simulation Evidence

5.1. Simulation Description

To examine the performance of our estimation methods, we consider the population models

$$\text{M1: } \text{logit}\{\text{pr}(\tilde{Y}_i=1|x_{i1})\}=\beta_0+\beta_1x_{i1}; \boldsymbol{\beta}=(\beta_0, \beta_1)' = (-3, 2)'$$

$$\text{M2: } \text{logit}\{\text{pr}(\tilde{Y}_i=1|x_{i1}, x_{i2})\}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}; \boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2)' = (-3, 2, -1)'$$

$$\text{M3: } \text{logit}\{\text{pr}(\tilde{Y}_i=1|x_{i1})\}=\beta_0+\beta_1x_{i1}+\beta_2x_{i1}^2; \boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2)' = (-3, 2, -0.5)',$$

where $x_{i1} \sim \mathcal{N}(0, 1)$ and $x_{i2} \sim \text{Bernoulli}(0.5)$. These settings give an overall disease prevalence of 13%, 10%, and 8% under M1, M2, and M3, respectively. To generate the individual true statuses, we first generated individual disease probabilities p_i according to each model and then simulated $\tilde{Y}_i \sim \text{Bernoulli}(p_i)$, $i = 1, 2, \dots, N$. This provides us with $\{(\tilde{Y}_i, \mathbf{x}_i'), i=1, 2, \dots, N\}$ for each model. Under each simulation setting, we repeat this process $B = 500$ times.

We consider three group testing protocols: master pool testing (MPT), Dorfman testing (DT), and two-dimensional array testing (AT). Recall that with MPT, non-overlapping pools are tested initially, but no further testing is performed; i.e., MPT is for estimation-only purposes. DT is a two-stage hierarchical procedure where non-overlapping pools are tested in the first stage (like MPT). Pools that test positively are resolved by retesting each individual separately in the second stage. AT is a two-stage non-hierarchical procedure where individuals are assigned to the cells of an array. In the first stage, the rows and columns of the array are tested. The second stage involves testing individuals who are not declared to be negative after seeing the row and column results. Because row/column results are potentially incorrect (due to assay error), we use the convention adopted in Kim et al. (2007) to determine which individuals are retested in the second stage with AT. For MPT and DT, we used master pools of size $c = 5$. For AT, we used master arrays of size 5×5 . Under MPT, DT, and AT, all testing responses Z_j were simulated from

$Z_j | \tilde{Z}_j \sim \text{Bernoulli}\{S_{e_j}\tilde{Z}_j + (1 - S_{p_j})(1 - \tilde{Z}_j)\}$, where $\tilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \tilde{Y}_i > 0)$ is the true status of the j th pool. For all three group testing protocols, we assigned individuals to pools at random.

For each model (M1, M2, and M3) and protocol (MPT, DT, and AT), we examine the performance of our methods under three settings for the assay accuracy probabilities. In the first, we assume $S_{e_j} = S_e$ and $S_{p_j} = S_p$, for $j = 1, 2, \dots, J$, where S_e and S_p are known. In the second (for DT/AT only), we make the same assumptions as in the first except that S_e and S_p are unknown and, thus, are estimated simultaneously with the regression parameter $\boldsymbol{\beta}$. In the third (for DT/AT only), we consider the possibility of using multiple assays. For both DT and AT, a screening assay is used on the master pools with accuracies $S_{e(1)}$ and $S_{p(1)}$. For individual retests with DT and AT, a confirmatory assay is used with accuracies $S_{e(2)}$ and $S_{p(2)}$. The $L = 2$ pairs of accuracies are assumed to be unknown and are estimated simultaneously with $\boldsymbol{\beta}$. Note that for the second setting, we found that MPT does not consistently provide enough information to identify both S_e and S_p and the regression parameter $\boldsymbol{\beta}$. This is not surprising as similar challenges arise even with individual testing (Jones et al., 2010).

5.2. Simulation Results

Our simulation results for the first, second, and third assay accuracy settings are presented in Tables 1, 2, and 3, respectively. For the results presented herein, we used $N = 5000$ individuals and flat priors on *all* unknown parameters; i.e., $\pi(\boldsymbol{\beta}) \propto 1$ and $S_{e(j)}, S_{p(j)} \sim \text{beta}(1, 1)$ independently. All results are based on $B = 500$ group testing data sets, each with 5000 posterior draws for each parameter sampled after a burn-in of 1000 draws. Standard MCMC diagnostics were used to assess convergence throughout and revealed no concerns.

Table 1 displays the results from estimating $\boldsymbol{\beta}$ when $S_e = 0.95$ and $S_p = 0.98$, both known and common for testing pools and individuals. For each of the $B = 500$ data sets, we estimated the posterior mean of each parameter. Values of “Bias” and “SSD” in Table 1 are the empirical bias and sample standard deviation of the 500 posterior mean estimates. We also calculated the average estimated posterior standard deviation, denoted by “ESE,” and consistently observed close agreement between SSD and ESE. Finally, values of “CP95” in Table 1 report the empirical coverage probabilities for nominal 95% equal-tail credible intervals. With 500 data sets, the margin of error associated with these probabilities, assuming a 99% confidence level, is 0.03. Therefore, even when no information is provided in the prior for $\boldsymbol{\beta}$, credible intervals attain their nominal level for each group testing protocol.

For comparison purposes in Table 1, we also estimated $\boldsymbol{\beta}$ under individual testing (IT) while assuming $S_e = 0.95$ and $S_p = 0.98$. Not surprisingly, MPT produces posterior mean estimates that have higher bias and are more variable than that for IT; with master pools of size $c = 5$, MPT uses only one-fifth the number of tests. However, both case identification protocols (DT and AT) provide estimates that generally have smaller bias and increased precision when compared to IT despite requiring far fewer tests. For example, under model M3 where the prevalence is about 8%, the improved estimates under DT and AT arise from about 3000 tests on average—a 40% reduction when compared to IT. Even though this “get more for less” phenomenon has been recently reported elsewhere in the regression literature with non-Bayesian methods (Zhang et al., 2013), we still find it to be rather remarkable.

Tables 2 and 3 summarize our results when the assay accuracy probabilities are unknown. There are two additional parameters to estimate in Table 2, namely, a common S_e and S_p for pools and individuals. In Table 3, there are four additional parameters, screening accuracies $S_{e(1)}$ and $S_{p(1)}$ for pools and confirmatory accuracies $S_{e(2)}$ and $S_{p(2)}$ for individuals. We make two observations. First, under both DT and AT, estimates of the accuracies in both tables exhibit little (if any) average bias, there is close agreement between SSD and ESE, and credible intervals attain their nominal level. In other words, even when providing no information in the prior distributions, our methods provide reliable inference for the assay accuracies. Second, when comparing the results for $\boldsymbol{\beta}$ in Tables 2 and 3 with those in Table 1, note that having to additionally estimate unknown assay accuracies does not have a large effect on estimating $\boldsymbol{\beta}$. Posterior mean estimates for $\boldsymbol{\beta}$ in Tables 2 and 3 remain largely unchanged on average when compared to those in Table 1 and are only slightly more variable.

5.3. Additional Simulations

We have performed four additional simulation studies that complement and reinforce the findings in this section; see Web Appendix C for complete details and discussion. First, we performed a simulation study using the informative prior choices for β from Section 3.2. This study is much smaller in size, using only $N = 500$ individuals, because prior distributions for β were chosen to be informative. Second, the results in Tables 1–3 summarize the performance of our regression methods when individuals are assigned to master pools at random, thereby allowing for covariate values that are heterogeneous within the pools. In Web Appendix C, we provide these same tables when individuals are assigned to pools homogeneously (i.e., pooling individuals with “similar” covariates). Third, for both DT and AT, we have examined the importance of the conditional independence assumption stated in Section 2; i.e., that testing results for master pools and individual retests are independent conditional on the individual true statuses. This study reveals that even a severe violation of this assumption does not degrade the overall performance of our estimation methods. Fourth, we use two years of individual chlamydia and gonorrhea testing data from Nebraska to simulate DT data analyses under a variety of prior models for β and the assay accuracies. This study complements our data analysis in Section 6, which uses real DT results from the neighboring state of Iowa.

6. Iowa Chlamydia Data

As part of federally sponsored sexually transmitted disease assessment and prevention programs, the State Hygienic Laboratory (SHL) at the University of Iowa tests thousands of Iowa residents for chlamydia and gonorrhea each year. In the midst of budgetary concerns in the late 1990s, along with the advent of nucleic acid amplification technology, the SHL switched from individual testing to group testing. This decision has proven to be cost-effective. At a per-test cost of about \$37, the SHL has enjoyed enormous savings from using group testing—approximately \$3.1 million during 2009–2014.

In recent years, we have collaborated with public health officials and lab technicians at the SHL on their use of group testing. To illustrate the regression methodology in this article, we analyze chlamydia testing data collected on $N = 13862$ female subjects during the 2014 calendar year. The SHL receives both endocervical swab and urine specimens each day from testing clinics located throughout the state. Screening procedures at the SHL call for swab specimens to be tested in pools (usually of size four) and urine specimens to be tested individually. Swab master pools are formed by combining individual specimens chronologically as they arrive at the SHL, and positive swab pools are resolved by using Dorfman testing (DT). The SHL uses the Aptima Combo 2 Assay (AC2A; Gen-Probe, San Diego) to test both swab and urine specimens. For swabs, the AC2A is used to test both pools and individuals. Pilot data describing the accuracy of the AC2A for individual testing are summarized in its product insert, available at www.fda.com; see also Gaydos et al. (2003). We also summarize these pilot data in Web Appendix C.

Working with the lead lab technician at the SHL, we have constructed a data set which consists of testing results for 2273 swab master pools of size $c = 4$, 12 swab master pools of size $c = 3$, one swab master pool of size $c = 2$, 416 individual swab specimens, and 4316

individual urine specimens. Dorfman retesting results on positive swab master pools are also included. To relate an individual's chlamydia disease status to the available covariates, we consider the logistic regression model

$$\text{logit}\{\text{pr}(\tilde{Y}_i=1|\mathbf{x}_i)\}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}, \quad (5)$$

for $i = 1, 2, \dots, 13862$, where x_1 denotes age (in years). The other covariates are binary; $x_2 = 1$ if the individual is Caucasian (0, otherwise), $x_3 = 1$ if a new sexual partner was reported in the last 90 days, $x_4 = 1$ if multiple partners were reported in the last 90 days, $x_5 = 1$ if the individual had contact with a partner having any STD reported in the previous year, and $x_6 = 1$ if the individual showed symptoms of infection (e.g., painful intercourse, etc.). In our analysis, we envision three sets of assay accuracy probabilities: $S_{\alpha(1)}$ and $S_{\rho(1)}$ for swab specimens tested in pools, $S_{\alpha(2)}$ and $S_{\rho(2)}$ for swab specimens tested individually, and $S_{\alpha(3)}$ and $S_{\rho(3)}$ for urine specimens tested individually. Therefore, along with the 7 regression parameters in Equation (5), there are a total of 13 parameters to estimate with the available data. We chose flat prior distributions for all parameters; i.e., $\pi(\boldsymbol{\beta}) \propto 1$ and $S_{\alpha(l)}, S_{\rho(l)} \sim \text{beta}(1, 1)$, for $l = 1, 2, 3$, and used 40000 posterior draws for each one sampled after a burn-in of 1000 draws. Fitting the model took about 7 minutes on a computer with a 2.9 GHz processor and 8 GB of memory and convergence was closely monitored.

Table 4 shows the posterior mean, the sample standard deviation of the posterior draws (ESE), and 95% equal-tail credible intervals for each parameter. Each of the covariate effects is strongly significant and is in the direction that one would expect based on epidemiological findings from observational studies of chlamydia infection in females (see, e.g., Navarro et al., 2002; Einwalter et al., 2005). Comparing the pool and individual estimates of assay sensitivity for swabs (i.e., of $S_{\alpha(1)}$ and $S_{\alpha(2)}$), there appears to be a mild dilution effect, which is not unexpected based on the experience of the lead lab technician at the SHL. The relatively lower and noisier estimate of $S_{\alpha(3)}$, the assay sensitivity for urine, is also expected because urine specimens are tested individually and are not retested. Finally, we also estimated the model in Equation (5) using very informative beta priors for $S_{\alpha(2)}$, $S_{\rho(2)}$, $S_{\alpha(3)}$, and $S_{\rho(3)}$ based on the individual AC2A pilot data in Web Appendix C. The results were nearly identical to those in Table 4 which assume flat priors for these parameters.

7. Discussion

We have developed a general Bayesian regression framework for group testing data with individually measured covariates. The novelty of our approach is that, by requiring only information on pool membership, it can be used with data collected from any group testing protocol. Furthermore, our framework allows one to simultaneously estimate the accuracy probabilities for any number of assays used during the screening process. Our simulation and data analysis sections reveal that our model fitting procedure is fast, flexible, and performs well even when little or no information is provided in the prior distributions. Additional studies in the Web-based Supplementary Materials provide examples of choosing informative prior distributions, examining the effects of pool composition, and investigating robustness issues in the light of our assumptions in Section 2.

We believe that different modeling extensions, such as the inclusion of random effects, covariate measurement error, spatial autocorrelation, or multivariate binary response (e.g., for multiple diseases), are all possible given the basic framework presented in this article. Because our general estimation approach attempts to recover true individual disease statuses, these extensions should perform similarly to the corresponding techniques developed for potentially misclassified individual testing data. In addition, for the methods presented in this article, or for any of these extensions, a reviewer has suggested that modeling assay accuracy probabilities directly as functions of the pool size may be worthwhile. This could be done by using the techniques outlined in McMahan et al. (2013) and Delaigle and Hall (2015) which exploit underlying assay biomarker distributions. These techniques do require additional assumptions but may also provide a more parsimonious approach when compared to eliciting several prior distributions for $S_{\alpha(j)}$ and $S_{\rho(j)}$; e.g., different sets of distributions for different pool sizes.

Finally, because the methods in this article can flexibly accommodate data from any group testing protocol, there is nothing to prevent one from re-estimating a model when new diagnoses are recorded; i.e., to update model estimates in “real time” and perhaps even continuously. This attractive feature of our work can be helpful in two ways. First, posterior estimates of the individual-level probabilities of disease can be used to design the “informative” protocols in McMahan, Tebbs, and Bilder (2012a, 2012b) and Black, Bilder, and Tebbs (2012, 2015). These protocols rely on having estimates of individual disease probabilities so that future pools can be constructed informatively. Second, these same estimates may also be useful in locating potentially misdiagnosed individuals. We are currently exploring this second idea for implementation with single-disease group testing protocols, with the goal of improving overall diagnostic accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Editor, the Associate Editor, and two anonymous referees for their helpful comments on an earlier version of this article. We also thank Lucy DesJardin, Jeffrey Benfer, and Kristofer Eveland at the State Hygienic Laboratory at the University of Iowa. This research was funded by Grant R01 AI121351 from the National Institutes of Health.

References

- Albert J, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 1993; 88:669–679.
- Basu S, Mukhopadhyay S. Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhya*. 2000; 62:372–387.
- Bedrick E, Christensen R, Johnson W. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*. 1996; 91:1450–1460.
- Black M, Bilder C, Tebbs J. Group testing in heterogeneous populations by using halving algorithms. *Journal of the Royal Statistical Society: Series C*. 2012; 61:277–290.
- Black M, Bilder C, Tebbs J. Optimal retesting configurations for hierarchical group testing. *Journal of the Royal Statistical Society: Series C*. 2015; 64:693–710.

- Chen P, Tebbs J, Bilder C. Group testing regression models with fixed and random effects. *Biometrics*. 2009; 65:1270–1278. [PubMed: 19210734]
- Delaigle A, Hall P. Nonparametric methods for group testing data, taking dilution into account. *Biometrika*. 2015; 102:871–887.
- Delaigle A, Hall P, Wishart J. New approaches to non- and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*. 2014; 101:567–585.
- Delaigle A, Meister A. Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*. 2011; 106:640–650.
- Dhand N, Johnson W, Toribio J. A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *Journal of Agricultural, Biological, and Environmental Statistics*. 2010; 15:452–473.
- Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–440.
- Einwalter L, Ritchie J, Ault K, Smith E. Gonorrhea and chlamydia infection among women visiting family planning clinics: Racial variation in prevalence and predictors. *Perspectives on Sexual and Reproductive Health*. 2005; 37:135–140. [PubMed: 16150661]
- Farrington C. Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine*. 1992; 11:1591–1597. [PubMed: 1359621]
- Gaydos C, Quinn T, Willis D, Weissfeld A, Hook E, Martin D, Ferrero D, Schachter J. 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*. 2003; 41:304–309. [PubMed: 12517865]
- Gamerman D. Sampling from the posterior distribution in generalized linear models. *Statistics and Computing*. 1997; 7:57–68.
- Gastwirth J, Johnson W. Screening with cost effective quality control: Potential applications to HIV and drug testing. *Journal of the American Statistical Association*. 1994; 89:972–981.
- Hanson T, Branscum A, Johnson W. Informative g-priors for logistic regression. *Bayesian Analysis*. 2014; 9:597–612.
- Heffernan A, Aylward L, Toms L, Sly P, Macleod M, Mueller J. Pooled biological specimens for human biomonitoring of environmental chemicals: Opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology*. 2014; 24:225–232. [PubMed: 24192659]
- Huang X. An improved test of latent-variable model misspecification in structural measurement error models for group testing data. *Statistics in Medicine*. 2009; 28:3316–3327. [PubMed: 19691036]
- Ibrahim J, Chen M. Power prior distributions for regression models. *Statistical Science*. 2000; 15:46–60.
- Jones G, Johnson W, Hanson T, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*. 2010; 66:855–863. [PubMed: 19764953]
- Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C. Comparison of group testing algorithms for case identification in the presence of testing error. *Biometrics*. 2007; 63:1152–1163. [PubMed: 17501946]
- Kleinman S, Strong D, Tegtmeier G, Holland P, Gorlin J, Cousins C, Chiacchierini R, Pietrelli L. Hepatitis B virus (HBV) DNA screening of blood donations in minipools with the COBAS AmpliScreen HBV test. *Transfusion*. 2005; 45:1247–1257. [PubMed: 16078909]
- Krajden M, Cook D, Mak A, Chu K, Chahil N, Steinberg M, Rekart M, Gilbert M. 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*. 2014; 61:132–137. [PubMed: 25037533]
- Lewis J, Lockary V, Kobic S. Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases*. 2012; 39:46–48. [PubMed: 22183846]
- Liu A, Liu C, Zhang Z, Albert P. Optimality of group testing in the presence of misclassification. *Biometrika*. 2012; 99:245–251. [PubMed: 23049137]
- McCullagh, P., Nelder, J. *Generalized Linear Models*. Chapman and Hall; London: 1989.

- McMahan C, Tebbs J, Bilder C. Informative Dorfman screening. *Biometrics*. 2012a; 68:287–296. [PubMed: 21762119]
- McMahan C, Tebbs J, Bilder C. Two-dimensional informative array testing. *Biometrics*. 2012b; 68:793–804. [PubMed: 22212007]
- McMahan C, Tebbs J, Bilder C. Regression models for group testing data with pool dilution effects. *Biostatistics*. 2013; 14:284–298. [PubMed: 23197382]
- Navarro C, Jolly A, Nair R, Chen Y. Risk factors for genital chlamydial infection. *Canadian Journal of Infectious Diseases*. 2002; 13:195–207. [PubMed: 18159391]
- Sarov B, Novack L, Beer N, Safi J, Soliman H, Pliskin J, Litvak E, Yaari A, Shinar E. Feasibility and cost-benefit of implementing pooled screening for HCVAg in small blood bank settings. *Transfusion Medicine*. 2007; 17:479–487. [PubMed: 17727618]
- Speybroeck N, Williams C, Lafia K, Devleeschauwer B, Berkvens D. Estimating the prevalence of infections in vector populations using pools of samples. *Medical and Veterinary Entomology*. 2012; 26:361–371. [PubMed: 22486773]
- Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. [PubMed: 11129470]
- Xie M. Regression analysis of group testing samples. *Statistics in Medicine*. 2001; 20:1957–1969. [PubMed: 11427952]
- Zhang B, Bilder C, Tebbs J. Group testing regression model estimation when case identification is a goal. *Biometrical Journal*. 2013; 55:173–189. [PubMed: 23401252]

Table 1

Known assay accuracy probabilities $S_e = 0.95$ and $S_d = 0.98$ (common for pools and individuals). Average bias of 500 posterior mean estimates (Bias), sample standard deviation of 500 posterior mean estimates (SSD), average of 500 estimates of the posterior standard deviation (ESE), and empirical coverage probability (CP95) of nominal 95% equal-tail credible intervals. For each model (M1, M2, and M3), the flat prior $\pi(\boldsymbol{\beta}) \propto 1$ is used for the regression coefficients. The total number of individuals is $N = 5000$. For Dorfman testing (DT) and array testing (AT), the percentage reduction in the average number of tests (when compared to individual testing) is shown in parentheses. MPT, DT, and AT use master pools of size $c = 5$.

Model	Parameter	Individual	MPT	DT	AT
M1	$\beta_0 = -3.00$	Bias (CP95)	-0.01 (0.95)	-0.02 (0.93)	-0.01 (0.96)
		SSD (ESE)	0.11 (0.11)	0.15 (0.15)	0.09 (0.09)
	$\beta_1 = 2.00$	Bias (CP95)	0.01 (0.95)	0.02 (0.93)	0.01 (0.96)
		SSD (ESE)	0.09 (0.10)	0.16 (0.15)	0.08 (0.09)
	Average number of tests	5000	1000	3429 (31%)	3370 (33%)
M2	$\beta_0 = -3.00$	Bias (CP95)	-0.03 (0.95)	-0.04 (0.95)	-0.02 (0.95)
		SSD (ESE)	0.13 (0.13)	0.17 (0.17)	0.11 (0.12)
	$\beta_1 = 2.00$	Bias (CP95)	0.02 (0.94)	0.04 (0.95)	0.02 (0.95)
		SSD (ESE)	0.10 (0.11)	0.15 (0.15)	0.09 (0.09)
	$\beta_2 = -1.00$	Bias (CP95)	-0.01 (0.95)	-0.03 (0.96)	-0.01 (0.96)
		SSD (ESE)	0.14 (0.14)	0.21 (0.22)	0.13 (0.13)
	Average number of tests	5000	1000	2679 (46%)	2787 (44%)
M3	$\beta_0 = -3.00$	Bias (CP95)	-0.04 (0.95)	-0.06 (0.95)	-0.02 (0.95)
		SSD (ESE)	0.13 (0.13)	0.19 (0.20)	0.10 (0.10)
	$\beta_1 = 2.00$	Bias (CP95)	0.08 (0.94)	0.15 (0.95)	0.04 (0.94)
		SSD (ESE)	0.23 (0.24)	0.44 (0.42)	0.17 (0.18)
	$\beta_2 = -0.50$	Bias (CP95)	-0.04 (0.93)	-0.07 (0.93)	-0.02 (0.93)
		SSD (ESE)	0.10 (0.10)	0.20 (0.19)	0.08 (0.08)
	Average number of tests	5000	1000	2981 (40%)	2999 (40%)

Table 2

Unknown assay accuracy probabilities (common for pools and individuals). Average bias of 500 posterior mean estimates (Bias), sample standard deviation of 500 posterior mean estimates (SSD), average of 500 estimates of the posterior standard deviation (ESE), and empirical coverage probability (CP95) of nominal 95% equal-tail credible intervals. For each model (M1, M2, and M3), the flat prior $\pi(\boldsymbol{\beta}) \propto 1$ is used for the regression coefficients and beta(1, 1) priors are used for S_e and S_p . The total number of individuals is $N = 5000$. DT and AT use master pools of size $c = 5$.

Model		$\beta_0 = -3.00$	$\beta_1 = 2.00$	$S_e = 0.95$	$S_p = 0.98$			
M1	DT	Bias (CP95)	-0.02 (0.94)	0.02 (0.92)	0.00 (0.94)	0.00 (0.94)		
		SSD (ESE)	0.11 (0.11)	0.10 (0.10)	0.01 (0.01)	0.01 (0.01)		
	AT	Bias (CP95)	-0.01 (0.94)	0.01 (0.93)	0.00 (0.96)	0.00 (0.95)		
		SSD (ESE)	0.10 (0.10)	0.09 (0.09)	0.01 (0.01)	0.01 (0.01)		
M2	DT	Bias (CP95)	-0.01 (0.96)	0.01 (0.96)	0.00 (0.94)	0.00 (0.94)		
		SSD (ESE)	0.12 (0.13)	0.10 (0.10)	0.13 (0.13)	0.02 (0.02)	0.01 (0.01)	
	AT	Bias (CP95)	-0.02 (0.94)	0.01 (0.96)	0.00 (0.95)	0.00 (0.95)	0.00 (0.94)	
		SSD (ESE)	0.12 (0.12)	0.09 (0.09)	0.13 (0.13)	0.01 (0.01)	0.01 (0.01)	
M3	DT	Bias (CP95)	-0.01 (0.95)	0.03 (0.96)	$\beta_2 = -0.50$	$S_e = 0.95$	$S_p = 0.98$	
		SSD (ESE)	0.12 (0.13)	0.20 (0.19)	0.09 (0.09)	-0.01 (0.95)	0.00 (0.98)	0.00 (0.95)
	AT	Bias (CP95)	-0.03 (0.96)	0.04 (0.94)	-0.01 (0.94)	0.02 (0.02)	0.01 (0.01)	0.00 (0.94)
		SSD (ESE)	0.11 (0.11)	0.21 (0.18)	0.10 (0.08)	0.00 (0.93)	0.01 (0.01)	0.01 (0.01)

Table 3

Unknown assay accuracy probabilities (two different assays). Average bias of 500 posterior mean estimates (Bias), sample standard deviation of 500 posterior mean estimates (SSD), average of 500 estimates of the posterior standard deviation (ESE), and empirical coverage probability (CP95) of nominal 95% equal-tail credible intervals. For each model (M1, M2, and M3), the flat prior $\pi(\beta) \propto 1$ is used for the regression coefficients and beta(1, 1) priors are used for $S_{\epsilon(1)}$, $S_{\mu(1)}$, $S_{\epsilon(2)}$, and $S_{\mu(2)}$. The total number of individuals is $N = 5000$. DT and AT use master pools of size $c = 5$.

Model	$\beta_0 = -3.00$	$\beta_1 = 2.00$	$\beta_2 = -1.00$	$S_{\epsilon(1)} = 0.95$	$S_{\mu(1)} = 0.98$	$S_{\epsilon(2)} = 0.98$	$S_{\mu(2)} = 0.99$
M1	Bias (CP95)	-0.01 (0.97)	0.03 (0.94)	-0.01 (0.97)	0.00 (0.98)	0.00 (0.98)	0.00 (0.97)
	DT						
	SSD (ESE)	0.10 (0.11)	0.10 (0.09)	0.02 (0.03)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Bias (CP95)	-0.02 (0.97)	0.01 (0.95)	0.00 (0.95)	0.00 (0.95)	0.00 (0.97)	0.00 (0.96)
M2	AT						
	SSD (ESE)	0.09 (0.10)	0.09 (0.09)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Bias (CP95)	0.00 (0.94)	0.04 (0.94)	-0.02 (0.97)	0.00 (0.96)	-0.01 (0.97)	0.00 (0.95)
	DT						
M3	SSD (ESE)	0.12 (0.13)	0.11 (0.10)	0.03 (0.04)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Bias (CP95)	-0.02 (0.96)	0.02 (0.93)	0.00 (0.95)	0.00 (0.96)	0.00 (0.96)	0.00 (0.95)
	AT						
	SSD (ESE)	0.11 (0.11)	0.10 (0.10)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Model	$\beta_0 = -3.00$	$\beta_1 = 2.00$	$\beta_2 = -0.50$	$S_{\epsilon(1)} = 0.95$	$S_{\mu(1)} = 0.98$	$S_{\epsilon(2)} = 0.98$	$S_{\mu(2)} = 0.99$
M1	Bias (CP95)	0.06 (0.95)	0.05 (0.94)	-0.06 (0.97)	0.00 (0.99)	-0.01 (0.98)	0.00 (0.97)
	DT						
	SSD (ESE)	0.12 (0.14)	0.19 (0.19)	0.05 (0.06)	0.01 (0.01)	0.01 (0.02)	0.01 (0.01)
	Bias (CP95)	-0.02 (0.96)	0.05 (0.95)	0.00 (0.96)	0.00 (0.96)	0.00 (0.98)	-0.01 (0.95)
M2	AT						
	SSD (ESE)	0.10 (0.10)	0.18 (0.18)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
	Bias (CP95)	0.00 (0.94)	0.05 (0.94)	-0.02 (0.94)	0.00 (0.99)	-0.01 (0.98)	0.00 (0.97)
	DT						
M3	SSD (ESE)	0.12 (0.14)	0.19 (0.19)	0.05 (0.06)	0.01 (0.01)	0.01 (0.02)	0.01 (0.01)
	Bias (CP95)	-0.02 (0.96)	0.05 (0.95)	-0.02 (0.94)	0.00 (0.96)	0.00 (0.98)	-0.01 (0.95)
	AT						
	SSD (ESE)	0.10 (0.10)	0.18 (0.18)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)

Table 4

Iowa chlamydia data. Results from estimating the model in Equation (5) with $\pi(\boldsymbol{\beta}) \propto 1$ and $S_{\alpha(l)}, S_{p(l)} \sim \text{beta}(1, 1)$, for $l = 1, 2, 3$. Posterior mean estimates, estimated posterior standard deviations (ESE), and 95% equal-tail credible intervals (CI) are provided.

Parameter	Description	Estimate	ESE	95% CI
β_0		-0.759	0.194	(-1.126, 0.368)
β_1	Age	-0.071	0.007	(-0.085, -0.058)
β_2	Race	-0.348	0.081	(-0.505, -0.190)
β_3	New partner	0.276	0.070	(0.139, 0.414)
β_4	Multiple partners	0.330	0.094	(0.144, 0.513)
β_5	Contact with STD	1.408	0.112	(1.189, 1.628)
β_6	Symptoms	0.290	0.077	(0.138, 0.439)
$S_{\alpha(1)}$	Swab pool	0.891	0.069	(0.742, 0.995)
$S_{\alpha(2)}$	Swab individual	0.998	0.002	(0.993, 1.000)
$S_{\alpha(3)}$	Urine individual	0.836	0.091	(0.646, 0.987)
$S_{p(1)}$	Swab pool	0.999	0.001	(0.997, 1.000)
$S_{p(2)}$	Swab individual	0.978	0.007	(0.964, 0.993)
$S_{p(3)}$	Urine individual	0.989	0.007	(0.974, 0.999)