Biostatistics (2020) **0**, 0, *pp.* 1–17
doi:10.1093/biostatistics/kxaa003

Generalized additive regression for group testing data

YAN LIU

School of Community Health Sciences, University of Nevada, Reno, 1664 N. Virginia St, Reno, NV 89557, USA

CHRISTOPHER S. MCMAHAN

School of Mathematical and Statistical Sciences, Clemson University, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA

JOSHUA M. TEBBS*

Department of Statistics, University of South Carolina, 1523 Greene St, Columbia, SC 29208, USA

tebbs@stat.sc.edu

COLIN M. GALLAGHER

School of Mathematical and Statistical Sciences, Clemson University, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA

CHRISTOPHER R. BILDER

Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall North, Lincoln, NE 68583, USA

Summary

In screening applications involving low-prevalence diseases, pooling specimens (e.g., urine, blood, swabs, etc.) through group testing can be far more cost effective than testing specimens individually. Estimation is a common goal in such applications and typically involves modeling the probability of disease as a function of available covariates. In recent years, several authors have developed regression methods to accommodate the complex structure of group testing data but often under the assumption that covariate effects are linear. Although linearity is a reasonable assumption in some applications, it can lead to model misspecification and biased inference in others. To offer a more flexible framework, we propose a Bayesian generalized additive regression approach to model the individual-level probability of disease with potentially misclassified group testing data. Our approach can be used to analyze data arising from any group testing protocol with the goal of estimating multiple unknown smooth functions of covariates, standard linear effects for other covariates, and assay classification accuracy probabilities. We illustrate the methods in this article using group testing data on chlamydia infection in Iowa.

Keywords: Bayesian regression; Binary regression; Gaussian process; Gaussian predictive process; Pooled testing; Specimen pooling.

1. INTRODUCTION

When screening individuals for sexually transmitted diseases, group testing can provide substantial cost savings when compared to testing specimens individually. The origin of group testing can be traced back to Dorfman (1943), who proposed it to screen military inductees for syphilis during World War II. This seminal work suggested that several individual blood specimens could be combined into a pool that would then be tested. If the pool tested negatively, each contributing individual could be diagnosed as negative at the expense of a single test. On the other hand, positive pools would be resolved by testing its individual specimens one by one. Although simple and intuitive, Dorfman's idea to pool individual specimens has had a profound impact on current disease screening practices. Group testing is used in applications involving a multitude of infectious diseases, including HIV (Westreich *and others*, 2008; Krajden *and others*, 2014), chlamydia and gonorrhea (Lewis *and others*, 2012), influenza (Van *and others*, 2012), and the Zika virus (Saá *and others*, 2018). The benefits of pooling have also been realized in other applications such as environmental testing (Heffernan *and others*, 2014), genetic association studies (Shi *and others*, 2014), and disease surveillance in animals (Dhand *and others*, 2010).

Group testing has become the focus of a great deal of statistical research over the last 75 years. This research generally focuses on either the so-called "case identification" or "estimation" problems. The former refers to the development, characterization, and optimization of group testing strategies for classification purposes; see Kim *and others* (2007). The latter, which is the focus of this article, involves using grouped data to estimate quantities characterizing a population of individuals. Many authors have used group testing to estimate a population proportion; see Liu *and others* (2012) for a review. More recently, research in this area has shifted towards estimating regression models which relate individual-level covariates (e.g., age, race, presence of symptoms, etc.) to the testing responses observed from assaying pooled specimens. Notable works in the development of group testing regression methods include parametric approaches by Vansteelandt *and others* (2000), Huang and Tebbs (2009), and Chen *and others* (2009) as well as the semiparametric and nonparametric approaches in Delaigle and Meister (2011), Delaigle *and others* (2014), and Delaigle and Hall (2015). When viewed collectively, a limitation of these and other contributions is that the corresponding models are estimated by using only the testing responses on initially formed master pools; i.e., subsequent responses from retesting individuals in positive master pools are not utilized. As a result, this generally leads to regression estimators which are less efficient than their would-be individual testing (IT) counterparts.

A smaller collection of methods has attempted regression estimation while incorporating master pool and retesting responses from group testing protocols. Xie (2001) and Zhang *and others* (2013) accomplished this for specific protocols (e.g., Dorfman testing [DT], higher-stage hierarchical testing, array testing [AT], etc.) using parametric inference via the expectation–maximization algorithm, while Wang *and others* (2014) developed single-index regression methods to incorporate additional retesting responses. Most recently, McMahan *and others* (2017) proposed a Bayesian approach for regression analysis of group testing data within a generalized linear model (GLM) framework. The strengths of this approach were 3-fold. First, this approach can incorporate retesting data from any case identification protocol in group testing; second, it can estimate assay accuracy probabilities along with the regression coefficients; and third, it can incorporate historical information about disease prevalence and assay performance.

In this article, we propose a Bayesian generalized additive modeling framework to estimate the probability of disease with group testing data. That is, the proposed approach utilizes a linear predictor which depends on unknown smooth functions of some covariates and linear combinations of others. This framework allows for a more flexible data analysis when compared to existing group testing regression methods which mandate linear covariate effects and therefore may simultaneously assuage concerns regarding

*To whom correspondence should be addressed.

model misspecification. Gaussian process (GP) (Rasmussen and Williams, 2006) and predictive process (GPP) (Banerjee *and others*, 2008) priors are employed to estimate the unknown functions, and our resulting modeling approach retains all of the strengths of McMahan *and others* (2017). To facilitate posterior estimation and inference, a computationally efficient sampling algorithm is constructed by introducing carefully structured latent random variables. We demonstrate that the resulting estimates are as accurate and as efficient as those that would have been obtained from analyzing individual-level testing responses within the same framework. This "get more for less" phenomenon makes group testing a cost-effective tool for disease screening in resource limited environments while producing equally good or better population-level estimates in the process.

The remainder of this article is organized as follows. In Section 2, we state modeling assumptions and present the proposed regression approach, complete with data augmentation steps and posterior sampling details, which can be used for any group testing protocol. In Section 3, we use simulation to assess the performance of our methods under a variety of settings for group testing protocols commonly used in practice. In Section 4, we analyze group testing data recently collected in Iowa as part of the state's surveillance program for chlamydia infection, illustrating the limitations of existing regression methods which assume linear covariate effects. In Section 5, we summarize our work and describe future areas of research.

2. Methodology

2.1. *Preliminaries*

Suppose group testing is used to test N individuals for a binary characteristic, such as disease status. In this article, our goal is to develop a regression framework which can be implemented for any group testing protocol. Let $\widetilde{Y}_i$ denote the true disease status of the ith individual, for $i = 1, ..., N$, that is, $\widetilde{Y}_i = 1$ if the ith individual is truly positive and $\widetilde{Y}_i = 0$ otherwise. Let $\mathbf{x}_i = (x_{i1}, ..., x_{iq})'$ denote a vector of covariates observed for the ith individual. We assume that individuals' true disease statuses are conditionally independent given the covariates and that the relationship between $\widetilde{Y}_i$ and $\mathbf{x}_i$ is given by the generalized additive model

$$H^{-1}\{\text{pr}(\widetilde{Y}_i = 1|\mathbf{x}_i)\} = \beta_0 + \sum_{l=1}^{q_1} g_l(x_{il}) + \sum_{l=1}^{q_2} \beta_l x_{i,q_1+l}, \tag{2.1}$$

where $H(\cdot)$ is a known binary link function, $\beta_l, l = 0, 1, ..., q_2$, are regression coefficients, $g_l(\cdot), l = 1, ..., q_1$, are unknown smooth functions, and $q_1 + q_2 = q$. We assume throughout that $H(\cdot)$ is the probit link, with generalizations to other link functions being straightforward after expressing $H(\cdot)$ as a Gaussian scale mixture; see, e.g., Albert and Chib (1993) and Polson *and others* (2013). Note that if the unknown functions $g_l(\cdot)$ are assumed to be linear, then the regression model in (2.1) would reduce to the GLM in McMahan *and others* (2017). Therefore, a primary focus of this work is to develop methods which can reliably estimate and draw inference on $g_l(\cdot)$ at any value in its support, which is denoted by $\mathcal{X}_l$, for $l = 1, ..., q_1$. To accomplish this within a Bayesian paradigm, both GP and GPP prior models are considered to obliquely represent $g_l(x)$, for $x \in \mathcal{X}_l$. These models are described in Section 2.3.

If the true disease statuses $\widetilde{Y}_i$ were observed, then the model in (2.1) could be estimated by using the methods in Choudhuri *and others* (2007). However, in most group testing scenarios, the $\widetilde{Y}_i$'s are never actually observed. This is true because individuals are pooled and also because most diagnostic assays do not have perfect sensitivity and specificity. Therefore, the process of deconvolving the observed group testing data back on to the individual level is error-laden and, depending on which retesting protocol is used, may involve numerous testing outcomes on the same individual tested in different pools. This unique

aspect is what makes the development of regression methodologies a nontrivial problem, especially those that can flexibly accommodate data arising from any group testing protocol.

To preserve this generality, as in McMahan *and others* (2017), let $\mathcal{P}_j, j = 1, ..., J$, denote the set of indices corresponding to those individuals assigned to the jth pool. Note that if IT was used, then $J = N$ and each $\mathcal{P}_j$ is a singleton; i.e., a "pool" of size one. Otherwise, our only requirement is that $\cup_j \mathcal{P}_j = \{1, \ldots, N\}$, allowing the methodology herein to be used for any group testing protocol in Kim *and others* (2007) or elsewhere. Let $\widetilde{Z}_j = 1$ if the jth pool is truly positive; i.e., $\widetilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0)$, where $I(\cdot)$ is the usual indicator function. In the presence of imperfect testing, the $\widetilde{Z}_j$'s, like the $\widetilde{Y}_i$'s, are never observed. Instead, the observed testing response for $\mathcal{P}_j$ is an error-contaminated version of $\widetilde{Z}_j$, which we denote by Z_j; that is, $Z_j = 1$ if the jth pool tests positively, $Z_j = 0$ otherwise. Note that $Z_j | \widetilde{Z}_j, S_{e_j}, S_{p_j} \sim \text{Bernoulli}\{S_{e_j}\widetilde{Z}_j + (1 - S_{p_j})(1 - \widetilde{Z}_j)\}$, where $S_{e_j} = \text{pr}(Z_j = 1 | \widetilde{Z}_j = 1)$ and $S_{p_j} = \text{pr}(Z_j = 0 | \widetilde{Z}_j = 0)$ denote the sensitivity and specificity, respectively, of the assay that tests the individual(s) in $\mathcal{P}_j$.

2.2. *Data augmentation*

Aggregate the N true individual disease statuses as $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, ..., \widetilde{Y}_N)'$ and the N covariate vectors as $\mathbf{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N)$. Define $g_{il} = g_l(x_{il})$, for $i = 1, \ldots, N$ and $l = 1, \ldots, q_1$, and let $\mathbf{Z} = (Z_1, \ldots, Z_J)'$, $\mathbf{S}_e = (S_{e_1}, \ldots, S_{e_J})'$, $\mathbf{S}_p = (S_{p_1}, \ldots, S_{p_J})'$, $\mathbf{g}_l = (g_{1l}, \ldots, g_{Nl})'$ for $l = 1, \ldots, q_1$, and $\mathbf{G} = (\mathbf{g}_1\ \mathbf{g}_2\ \cdots\ \mathbf{g}_{q_1})$. The observed data likelihood can be expressed as

$$\pi(\mathbf{Z}|\mathbf{S}_e, \mathbf{S}_p, \boldsymbol{\beta}, \mathbf{G}, \mathbf{X}) = \sum_{\widetilde{\mathbf{Y}} \in \{0,1\}^N} \left[\prod_{j=1}^{J} \{S_{e_j}^{Z_j}(1 - S_{e_j})^{1-Z_j}\}^{\widetilde{Z}_j} \{(1 - S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{1-\widetilde{Z}_j} \times \prod_{i=1}^{N} H(\eta_i)^{\widetilde{Y}_i} \{1 - H(\eta_i)\}^{1-\widetilde{Y}_i} \right], \tag{2.2}$$

where $\eta_i = \sum_{l=1}^{q_1} g_{il} + \mathbf{x}_{i2}'\boldsymbol{\beta}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q_2})'$ is a vector of regression coefficients corresponding to the covariates in $\mathbf{x}_{i2} = (1, x_{i,q_1+1}, \ldots, x_{iq})'$, for $i = 1, \ldots, N$. The set $\{0, 1\}^N$ denotes the collection of all 2^N possible realizations of $\widetilde{\mathbf{Y}}$. Note that in writing (2.2), we assume the observed testing responses in $\mathbf{Z}$ are conditionally independent given the true disease statuses in $\widetilde{\mathbf{Y}}$ and that the conditional distribution $\mathbf{Z}|\widetilde{\mathbf{Y}}$ does not depend on the covariates in $\mathbf{X}$. Simulation studies in McMahan *and others* (2017) reveal that mild to even moderate violations of these assumptions do not compromise estimation within a GLM setting.

Evaluating (2.2) is computationally infeasible when N is large; thus, developing a posterior sampling algorithm that requires such an evaluation is generally not possible. We propose a two-stage data augmentation procedure to facilitate the development of such an algorithm. The first stage introduces individuals' true disease statuses $\widetilde{Y}_i$ as latent random variables and the second exploits the fact that $H(\cdot)$ is the probit link. After the first stage, the joint conditional distribution of $\mathbf{Z}$ and $\widetilde{\mathbf{Y}}$ can be expressed as

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}|\mathbf{S}_e, \mathbf{S}_p, \boldsymbol{\beta}, \mathbf{G}, \mathbf{X}) = \prod_{j=1}^{J} \{S_{e_j}^{Z_j}(1 - S_{e_j})^{1-Z_j}\}^{\widetilde{Z}_j} \{(1 - S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{1-\widetilde{Z}_j} \times \prod_{i=1}^{N} H(\eta_i)^{\widetilde{Y}_i} \{1 - H(\eta_i)\}^{1-\widetilde{Y}_i}. \tag{2.3}$$

Following Albert and Chib (1993), the second stage introduces mutually independent normal random variables $U_i \sim \mathcal{N}(\eta_i, 1)$, for $i = 1, \ldots, N$, such that $U_i > 0$ if $\widetilde{Y}_i = 1$ and $U_i \leq 0$ if $\widetilde{Y}_i = 0$. This data

augmentation step yields

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \mathbf{U}|\mathbf{S}_e, \mathbf{S}_p, \boldsymbol{\beta}, \mathbf{G}, \mathbf{X}) \propto \prod_{j=1}^{J} \{S_{e_j}^{Z_j}(1 - S_{e_j})^{1-Z_j}\}^{\widetilde{Z}_j} \{(1 - S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{1-\widetilde{Z}_j}$$
$$\times \prod_{i=1}^{N} \zeta(U_i - \eta_i)\{I(\widetilde{Y}_i = 1, U_i > 0) + I(\widetilde{Y}_i = 0, U_i \leq 0)\}, \tag{2.4}$$

where $\zeta(\cdot)$ denotes the standard normal probability density function and $\mathbf{U} = (U_1, ..., U_N)'$.

2.3. *Modeling the unknown functions*

Our regression methods allow for GP and GPP prior models to represent the unknown functions $g_l(\cdot)$ in (2.1). Let $\{x^*_{1l}, \ldots, x^*_{K_l l}\}$ denote the K_l unique values among $\{x_{1l}, \ldots, x_{Nl}\}$ and define $\mathbf{g}^*_l = (g_l(x^*_{1l}), \ldots, g_l(x^*_{K_l l}))'$, for $l = 1, \ldots, q_1$. Under the GP model, the $\mathbf{g}^*_l$'s are mutually independent normal random vectors with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}^*_l = \tau_l^{-1}\mathbf{R}^*_l$, where τ_l is a precision parameter and the (k, k')th element of $\mathbf{R}^*_l$ is $\rho_l(x^*_{kl}, x^*_{k'l}; \boldsymbol{\theta}_l)$, where $\rho_l = \rho_l(\cdot, \cdot; \cdot)$ is a correlation function and $\boldsymbol{\theta}_l$ consists of smoothness and decay parameters. To relate $\mathbf{g}_l = (g_{1l}, ..., g_{Nl})'$ to its subvector $\mathbf{g}^*_l$, let $\mathbf{M}_l = (m^{(l)}_{kk'})$ be an $N \times K_l$ matrix, where $m^{(l)}_{kk'} = 1$ if $x_{kl} = x^*_{k'l}$ and $m^{(l)}_{kk'} = 0$ if $x_{kl} \neq x^*_{k'l}$ so that $\mathbf{g}_l = \mathbf{M}_l\mathbf{g}^*_l$. The linear predictor η_i under the GP model can be written as

$$\eta_i = \sum_{l=1}^{q_1} \mathbf{m}'_{il}\mathbf{g}^*_l + \mathbf{x}'_{i2}\boldsymbol{\beta},$$

where $\mathbf{m}'_{il}$ is the ith row of $\mathbf{M}_l$. Pairing this with (2.4), the full conditional distribution of $\mathbf{g}^*_l$ is normal with mean $\boldsymbol{\mu}^*_l$ and covariance matrix $\boldsymbol{\Sigma}^*_l$; see Section 2.4 for exact expressions. Note that drawing samples from this distribution requires one to calculate the determinant and inverse of $\boldsymbol{\Sigma}^*_l$, which could be a very large $(K_l \times K_l)$ dense matrix. Furthermore, the computational burden in working with $\boldsymbol{\Sigma}^*_l$ is only exacerbated when these calculations are needed within each step of an Markov chain Monte Carlo algorithm; i.e., if $\boldsymbol{\theta}_l$ is to be sampled along with the other model parameters.

To reduce this complexity, a GPP prior can be used instead. This approach specifies a "parent" process based on strategically chosen knots and then interpolates this process to the points of interest. Let $\{\widetilde{x}_{1l}, \ldots, \widetilde{x}_{\widetilde{K}_l l}\}$ denote the selected knots within the support $\mathcal{X}_l$, where $\widetilde{K}_l << K_l$, and let $\widetilde{\mathbf{g}}_l = (g_l(\widetilde{x}_{1l}), \ldots, g_l(\widetilde{x}_{\widetilde{K}_l l}))'$. For each $l = 1, \ldots, q_1$, the GP model yields $\widetilde{\mathbf{g}}_l|\tau_l, \boldsymbol{\theta}_l \overset{ind}{\sim} \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{C}}_l)$, where $\widetilde{\mathbf{C}}_l = \tau_l^{-1}\widetilde{\mathbf{R}}_l$ and the (k, k')th element of $\widetilde{\mathbf{R}}_l$ is $\rho_l(\widetilde{x}_{kl}, \widetilde{x}_{k'l}; \boldsymbol{\theta}_l)$. It follows that $\widetilde{\mathbf{g}}_l$ and $\mathbf{g}^*_l$ are jointly multivariate normal, that is,

$$\begin{pmatrix} \widetilde{\mathbf{g}}_l \\ \mathbf{g}^*_l \end{pmatrix} \Bigg| \tau_l, \boldsymbol{\theta}_l \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\tau_l}\begin{pmatrix} \widetilde{\mathbf{R}}_l & \widetilde{\mathbf{R}}^*_l \\ \widetilde{\mathbf{R}}^{*\prime}_l & \mathbf{R}^*_l \end{pmatrix}\right), \tag{2.5}$$

where $\mathbf{R}^*_l$ is defined above and $\widetilde{\mathbf{R}}^*_l$ is a $\widetilde{K}_l \times K_l$ matrix whose (k, k')th element is $\rho_l(\widetilde{x}_{kl}, x^*_{k'l}; \boldsymbol{\theta}_l)$. The GPP model exploits the relationship in (2.5) by replacing $\mathbf{g}^*_l$ in the GP model with its conditional expectation given $\widetilde{\mathbf{g}}_l$, which is $\mathbf{T}_l\widetilde{\mathbf{g}}_l$, where $\mathbf{T}_l = \widetilde{\mathbf{R}}^{*\prime}_l\widetilde{\mathbf{R}}^{-1}_l$. Therefore, the linear predictor η_i under the GPP model can

be represented as

$$\eta_i = \sum_{l=1}^{q_1} \mathbf{m}_{il}' \mathbf{T}_l \widetilde{\mathbf{g}}_l + \mathbf{x}_{i2}' \boldsymbol{\beta}.$$

When compared to a GP prior model, the computational burden associated with GPP is potentially much less because posterior sampling involves matrices of dimension $\widetilde{K}_l \times \widetilde{K}_l$.

The functions $g_l(\cdot)$, $l = 1, ..., q_1$, are identifiable up to a constant and therefore side conditions are needed for estimation. A common restriction requires $g_l(\cdot)$ to integrate to 0 over $\mathcal{X}_l$, $l = 1, \ldots, q_1$, which can be enforced in finite samples by requiring $\sum_{k=1}^{K_l} g_l(x_{kl}^*) = 0$ for GP and $\sum_{k=1}^{\widetilde{K}_l} g_l(\widetilde{x}_{kl}) = 0$ for GPP; see Friedman *and others* (2001). Under either the GP or GPP priors, one can exploit the relationship in (2.5) to interpolate a functional estimate of $g_l(x)$ for any $x \in \mathcal{X}_l$, with the usual cautions in place regarding extrapolation.

2.4. *Posterior sampling*

We describe posterior sampling for $\boldsymbol{\beta}$ and the unknown functions $g_l(\cdot)$, $l = 1, \ldots, q_1$, when the assay accuracy probabilities in $\mathbf{S}_e$ and $\mathbf{S}_p$ are known; Section 2.5 generalizes this by allowing $\mathbf{S}_e$ and $\mathbf{S}_p$ to be unknown. In what follows, $\mathbf{G} = (\mathbf{g}_1\ \mathbf{g}_2\ \cdots\ \mathbf{g}_{q_1})$ and $\mathbf{g}_l$ equals $\mathbf{M}_l \mathbf{g}_l^*$ or $\mathbf{M}_l \mathbf{T}_l \widetilde{\mathbf{g}}_l$, depending on whether a GP or GPP prior is used, respectively.

Let $\pi(\boldsymbol{\beta})$ denote a prior distribution for $\boldsymbol{\beta}$. From (2.4), the full conditional distribution $\pi(\boldsymbol{\beta}|\mathbf{U}, \mathbf{G})$ satisfies $\pi(\boldsymbol{\beta}|\mathbf{U}, \mathbf{G}) \propto \exp\{-(\mathbf{U} - \sum_{l=1}^{q_1} \mathbf{g}_l - \mathbf{X}_2\boldsymbol{\beta})'(\mathbf{U} - \sum_{l=1}^{q_1} \mathbf{g}_l - \mathbf{X}_2\boldsymbol{\beta})/2\}\pi(\boldsymbol{\beta})$, where $\mathbf{X}_2 = (\mathbf{x}_{12}\ \mathbf{x}_{22}\ \cdots\ \mathbf{x}_{N2})'$ is the design matrix associated with the linear covariate effects. Thus, it is natural to specify a $\mathcal{N}(\mathbf{a}, \boldsymbol{\Gamma})$ prior for $\boldsymbol{\beta}$, where the mean and covariance matrix hyperparameters can be chosen diffusely or informatively to incorporate historical data. This specification leads to

$$\boldsymbol{\beta}|\mathbf{U}, \mathbf{G} \ \sim\ \mathcal{N}\left((\mathbf{X}_2'\mathbf{X}_2 + \boldsymbol{\Gamma}^{-1})^{-1} \left\{ \boldsymbol{\Gamma}^{-1}\mathbf{a} + \mathbf{X}_2' \left(\mathbf{U} - \sum_{l=1}^{q_1} \mathbf{g}_l \right) \right\},\ (\mathbf{X}_2'\mathbf{X}_2 + \boldsymbol{\Gamma}^{-1})^{-1} \right).$$

Furthermore, it follows from (2.4) that the full conditional distribution of U_i is truncated normal, where the truncation depends on the ith latent disease status; i.e.,

$$U_i|\widetilde{Y}_i, \boldsymbol{\beta}, \mathbf{G} \ \sim\ \mathcal{TN}\{\eta_i, 1, (0, \infty)\} I(\widetilde{Y}_i = 1) + \mathcal{TN}\{\eta_i, 1, (-\infty, 0)\} I(\widetilde{Y}_i = 0),\ \ i = 1, \ldots, N,$$

where $\mathcal{TN}\{\mu, \sigma^2, (a, b)\}$ denotes a truncated normal distribution with mean μ, variance σ^2, and support (a, b). It also follows from (2.3) that the full conditional distribution of the ith disease status

$$\widetilde{Y}_i|\mathbf{Z}, \widetilde{\mathbf{Y}}_{-i}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{G}, \boldsymbol{\beta} \ \sim\ \text{Bernoulli}\left(\frac{p_{i1}^*}{p_{i1}^* + p_{i0}^*} \right),$$

where $\widetilde{\mathbf{Y}}_{-i} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_{i-1}, \widetilde{Y}_{i+1}, \ldots, \widetilde{Y}_N)'$, $p_{i1}^* = H(\eta_i) \prod_{j \in \mathcal{A}_i} S_{e_j}^{Z_j} (1 - S_{e_j})^{1 - Z_j}$,

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

$\mathcal{A}_i = \{j : i \in \mathcal{P}_j\}$, and $\mathcal{P}_{ij} = \{i' \in \mathcal{P}_j : i \neq i'\}$. Derivation details for this conditional distribution under a GLM are described in McMahan *and others* (2017).

We now turn our attention to updating the unknown functions. Under the GP model, the full conditional distribution of $\mathbf{g}_l^*$ can be expressed as $\pi(\mathbf{g}_l^*|\mathbf{U},\boldsymbol{\beta},\tau_l,\boldsymbol{\theta}_l,\mathbf{G}_{(-l)}) \propto \exp\{-(\mathbf{U}^* - \mathbf{M}_l\mathbf{g}_l^*)'(\mathbf{U}^* - \mathbf{M}_l\mathbf{g}_l^*)/2\}\exp\{\mathbf{g}_l^{*'}\mathbf{C}_l^{*-1}\mathbf{g}_l^*/2\}$, where $\mathbf{U}^* = \mathbf{U} - \mathbf{X}_2\boldsymbol{\beta} - \sum_{l'\neq l}\mathbf{M}_{l'}\mathbf{g}_{l'}^*$ and $\mathbf{G}_{(-l)}$ is the matrix $\mathbf{G}$ with $\mathbf{g}_l$ removed. Straightforward algebra yields $\mathbf{g}_l^*|\mathbf{U},\boldsymbol{\beta},\tau_l,\boldsymbol{\theta}_l,\mathbf{G}_{(-l)} \sim \mathcal{N}(\boldsymbol{\mu}_l^*,\boldsymbol{\Sigma}_l^*)$, where $\boldsymbol{\mu}_l^* = \boldsymbol{\Sigma}_l^*\mathbf{M}_l'(\mathbf{U}-\mathbf{X}_2\boldsymbol{\beta}-\sum_{l'\neq l}\mathbf{M}_{l'}\mathbf{g}_{l'}^*)$ and $\boldsymbol{\Sigma}_l^* = (\mathbf{M}_l'\mathbf{M}_l+\mathbf{C}_l^{*-1})^{-1}$. Under the GPP model, a similar argument shows the full conditional distribution $\widetilde{\mathbf{g}}_l|\mathbf{U},\boldsymbol{\beta},\tau_l,\boldsymbol{\theta}_l,\mathbf{G}_{(-l)} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_l,\widetilde{\boldsymbol{\Sigma}}_l)$, where $\widetilde{\boldsymbol{\mu}}_l = \widetilde{\boldsymbol{\Sigma}}_l\mathbf{T}_l'\mathbf{M}_l'(\mathbf{U}-\mathbf{X}_2\boldsymbol{\beta}-\sum_{l'\neq l}\mathbf{M}_{l'}\mathbf{T}_{l'}\widetilde{\mathbf{g}}_{l'})$ and $\widetilde{\boldsymbol{\Sigma}}_l = (\mathbf{T}_l'\mathbf{M}_l'\mathbf{M}_l\mathbf{T}_l+\widetilde{\mathbf{C}}_l^{-1})^{-1}$. Therefore, all that remains are the GP/GPP hyperparameters; i.e., the precision parameter τ_l and $\boldsymbol{\theta}_l$. Exploiting conditional conjugacy, independent gamma priors with shape a_l and rate b_l are used for τ_l, which lead to full conditional distributions $\tau_l|\mathbf{g}_l^*,\boldsymbol{\theta}_l \sim \text{gamma}(a_l+K_l/2, b_l+\mathbf{g}_l^{*'}\mathbf{R}_l^{*-1}\mathbf{g}_l^*/2)$ and $\tau_l|\widetilde{\mathbf{g}}_l,\boldsymbol{\theta}_l \sim \text{gamma}(a_l+\widetilde{K}_l/2, b_l+\widetilde{\mathbf{g}}_l'\widetilde{\mathbf{R}}_l^{-1}\widetilde{\mathbf{g}}_l/2)$ under the GP and GPP models, respectively. As for $\boldsymbol{\theta}_l$, its full conditionals satisfy $\pi(\boldsymbol{\theta}_l|\mathbf{g}_l^*,\tau_l) \propto |\mathbf{C}_l^*|^{-1/2}\exp\{-\mathbf{g}_l^{*'}\mathbf{C}_l^{*-1}\mathbf{g}_l^*/2\}\pi(\boldsymbol{\theta}_l)$ and $\pi(\boldsymbol{\theta}_l|\mathbf{U},\widetilde{\mathbf{g}}_l,\tau_l,\mathbf{X}) \propto |\widetilde{\mathbf{C}}_l|^{-1/2}\exp[-\{(\mathbf{U}-\boldsymbol{\eta})'(\mathbf{U}-\boldsymbol{\eta})+\widetilde{\mathbf{g}}_l'\widetilde{\mathbf{C}}_l^{-1}\widetilde{\mathbf{g}}_l\}/2]\pi(\boldsymbol{\theta}_l)$, respectively, where $\boldsymbol{\eta} = \sum_{l=1}^{q_1}\mathbf{M}_l\mathbf{T}_l\widetilde{\mathbf{g}}_l+\mathbf{X}_2\boldsymbol{\beta}$. Because $\boldsymbol{\theta}_l$ is implicitly a part of $\mathbf{C}_l^*$, $\widetilde{\mathbf{C}}_l$, and $\mathbf{T}_l$, determining a conjugate prior for $\boldsymbol{\theta}_l$ is not straightforward under any correlation function ρ_l. We therefore use a random walk Metropolis–Hastings algorithm to draw samples of $\boldsymbol{\theta}_l$, as outlined in the supplementary material available at *Biostatistics* online.

2.5. *Unknown assay accuracy probabilities*

The results in Section 2.4 describe the steps necessary to estimate the model in (2.1) when the assay accuracy probabilities in $\mathbf{S}_e = (S_{e_1},...,S_{e_J})'$ and $\mathbf{S}_p = (S_{p_1},\ldots,S_{p_J})'$ are known. However, allowing $\mathbf{S}_e$ and $\mathbf{S}_p$ to be unknown and estimating them simultaneously with $\boldsymbol{\beta}$ and $g_l(\cdot)$, $l = 1,\ldots,q_1$, is also possible. As in McMahan *and others* (2017), define $\mathcal{M}(m) = \{j : \text{the } m\text{th assay was used to test the } j\text{th pool}\}$, for $m = 1,\ldots,M$, so that $\mathbf{S}_e$ and $\mathbf{S}_p$ can be expressed as $\mathbf{S}_e = (S_{e(1)},\ldots,S_{e(M)})'$ and $\mathbf{S}_p = (S_{p(1)},\ldots,S_{p(M)})'$; i.e., this reparametrization requires one only to keep track of which assay was used to test which pool. For example, if a group testing protocol uses a highly specific screening assay to test pooled specimens and a confirmatory assay for individual specimens, then $M = 2$ and there are four probabilities to estimate. In other applications, one might want $S_{e(m)}$ and $S_{p(m)}$ to depend on which type of specimen is tested (e.g., urine, swab, etc.) or even the pool sizes used; see Section 4.

Due to the form of (2.3), we specify independent beta priors $S_{e(m)} \sim \text{beta}(a_{S_{e(m)}}, b_{S_{e(m)}})$ and $S_{p(m)} \sim \text{beta}(a_{S_{p(m)}}, b_{S_{p(m)}})$, for $m = 1,\ldots,M$. The full conditional distributions are $S_{e(m)}|\mathbf{Z},\widetilde{\mathbf{Y}} \sim \text{beta}(a^*_{S_{e(m)}}, b^*_{S_{e(m)}})$ and $S_{p(m)}|\mathbf{Z},\widetilde{\mathbf{Y}} \sim \text{beta}(a^*_{S_{p(m)}}, b^*_{S_{p(m)}})$, where $a^*_{S_{e(m)}} = a_{S_{e(m)}} + \sum_{j\in\mathcal{M}(m)} Z_j\widetilde{Z}_j$, $b^*_{S_{e(m)}} = b_{S_{e(m)}} + \sum_{j\in\mathcal{M}(m)}(1 - Z_j)\widetilde{Z}_j$, $a^*_{S_{p(m)}} = a_{S_{p(m)}} + \sum_{j\in\mathcal{M}(m)}(1 - Z_j)(1 - \widetilde{Z}_j)$, and $b^*_{S_{p(m)}} = b_{S_{p(m)}} + \sum_{j\in\mathcal{M}(m)} Z_j(1 - \widetilde{Z}_j)$. Pilot data on assay performance, which are typically available in the product literature published by manufacturers, can be used to construct informative prior distributions for $S_{e(m)}$ and $S_{p(m)}$. However, for group testing protocols which involve retesting individuals for case identification, we have observed these parameters can be estimated correctly even when one injects little or no information into the prior distributions; see Section 3.

The complete posterior sampling algorithm to estimate $\boldsymbol{\beta}$ and $g_l(\cdot)$, $l = 1,\ldots,q_1$, as well as the assay accuracy probabilities in $\mathbf{S}_e$ and $\mathbf{S}_p$ is given in the supplementary material available at *Biostatistics* online.

3. Simulation evidence

We consider two population-level models, both of which are of the form

$$H^{-1}\{\text{pr}(\widetilde{Y}_i = 1|\mathbf{x}_i)\} = \beta_0 + g_1(x_{i1}) + g_2(x_{i2}) + \beta_1 x_{i3} + \beta_2 x_{i4},$$

$i = 1, \ldots, N$, where $H(\cdot)$ is the probit link, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (-1.8, 0.5, 0.5)'$, $x_{i1}, x_{i2} \sim \mathcal{U}(-3, 3)$, $x_{i3} \sim \mathcal{N}(0, 1)$, and $x_{i4} \sim$ Bernoulli(0.5). In the first model (M1), the functions

$$g_1(x_1) = 0.7 \exp[\{-I(x_1 > 0)1.2^2 - I(x_1 < 0)1.2^{-2}\}x_1^2/6.25] - 0.468$$
$$g_2(x_2) = 0.6 \exp\left\{-\frac{(x_2 + 1.5)^2}{0.72}\right\} + 0.4 \exp\left\{-\frac{(x_2 - 1.5)^2}{1.28}\right\} - 0.279,$$

while in the second model (M2),

$$g_1(x_1) = \frac{0.2 \sin\{\pi(x_1 + 0.2)/2.5\} + 0.4}{\exp[\{-x_1 - (x_1 - 0.3)^2 I(x_1 > 0.3)\}/6]} - 0.351 \quad \text{and} \quad g_2(x_2) = \frac{4 \exp(1 + 1.5x_2)}{6 + 6 \exp(1 + 1.5x_2)} - 0.406.$$

These functions were chosen to cover a broad range of nonlinear patterns, and the additive constants (e.g., -0.468) were chosen to ensure the functions integrate to 0 over $\mathcal{X}_l = (-3, 3)$, $l = 1, 2$. Other parameter settings were selected to keep the population prevalence around 9%, which is consistent with our application in Section 4. The covariates $x_{i1}, x_{i2} \sim \mathcal{U}(-3, 3)$ were each rounded to the second decimal place so that we can compare the GP and GPP approaches. Doing this constrains the dimension of $\boldsymbol{\Sigma}_l^*$ to be at most 601×601, making GP computationally feasible. When estimating $g_1(\cdot)$ and $g_2(\cdot)$ under a GPP prior, we used 100 equally spaced knots within $(-3, 3)$; i.e., $\widetilde{K}_l = 100$, for $l = 1, 2$.

We generated $N = 5000$ individual true statuses $\widetilde{Y}_i$ from each regression model (M1 and M2) and randomly assigned these individuals to 1000 master pools, each of size five. As in McMahan *and others* (2017), we simulate the testing responses $\mathbf{Z}$ from three group testing protocols: master pool testing (MPT), DT, and two-dimensional AT. Briefly, MPT is for estimation only as positive pools are not resolved further. DT and AT are both two-stage protocols. Positive master pools in DT are resolved in the second stage by testing each individual, while AT arranges master pools in overlapping rows and columns in the first stage and uses IT in the second (Kim *and others*, 2007; McMahan *and others*, 2012). To incorporate the effect of imperfect testing, master pool responses in MPT, DT, and AT were simulated using $S_{e(1)} = 0.95$ and $S_{p(1)} = 0.98$, and individual retests for DT and AT were simulated using $S_{e(2)} = 0.98$ and $S_{p(2)} = 0.99$. This entire process was repeated $B = 500$ times for each group testing protocol (MPT, DT, and AT) and regression model (M1 and M2).

The following prior distributions were used. For the precision parameters in the covariance functions (for GP and GPP), we used $\tau_l \sim$ gamma$(a_l = 2, b_l = 1)$, for $l = 1, 2$, and the regression parameters β_0, β_1, and β_2 were each assigned vague $\mathcal{N}(0, 1000)$ priors. When assay accuracies were assumed to be unknown, we assigned uniform priors; i.e., $S_{e(m)}, S_{p(m)} \sim$ beta$(1, 1)$. Finally, for the GP and GPP priors, we used the flexible Matérn correlation function

$$\rho_l(x, x'; \boldsymbol{\theta}_l) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{|x - x'|}{\phi}\right)^{\nu} \kappa_\nu \left(\frac{|x - x'|}{\phi}\right), \tag{3.6}$$

where $\boldsymbol{\theta}_l = (\nu, \phi)'$, $l = 1, 2$, and $\kappa_\nu(\cdot)$ is the modified Bessel function of the third kind of order ν. In this submodel, ν controls the smoothness of the GP sample path and ϕ controls the decay rate. Following Banerjee *and others* (2008), we took $\nu = 2$ to obtain a desired degree of differentiability and assigned $\phi \sim \mathcal{U}(0.075, 0.750)$. When $\nu = 2$, the distance at which the correlation drops to 0.05 is approximately 6ϕ; thus, our prior for ϕ provides a range from 0.45 to 4.5 for this distance. The posterior sampling algorithm in the supplementary material available at *Biostatistics* online contains additional details on initial values. We used this algorithm to draw 5000 samples (after a burn-in of 2000 samples) and retained every 5th iterate. Trace plots were used to monitor convergence and consistently demonstrated excellent mixing.

Table 1. *Simulation results for models M1 and M2 under GP and GPP priors when assay accuracy probabilities are known. 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. Note that close agreement between SSD and ESE is preferred. The number of individuals is* $N = 5000$. *For DT and AT, the percentage reduction in the average number of tests (when compared to IT) is shown in parentheses. The average time (in minutes) to estimate the model is also shown.*

Model	Parameter		Individual	MPT	DT	AT
M1/GP	$\beta_1 = 0.50$	Bias (CP95)	0.01 (0.94)	0.01 (0.95)	0.01 (0.93)	0.01 (0.93)
		SSD (ESE)	0.03 (0.03)	0.07 (0.07)	0.03 (0.03)	0.03 (0.03)
	$\beta_2 = 0.50$	Bias (CP95)	0.01 (0.94)	0.02 (0.93)	0.01 (0.95)	0.01 (0.93)
		SSD (ESE)	0.06 (0.06)	0.14 (0.13)	0.06 (0.06)	0.06 (0.06)
	Time (in minutes)		284	291	284	286
M1/GPP	$\beta_1 = 0.50$	Bias (CP95)	0.01 (0.94)	0.02 (0.93)	0.01 (0.94)	0.01 (0.93)
		SSD (ESE)	0.03 (0.03)	0.07 (0.07)	0.03 (0.03)	0.03 (0.03)
	$\beta_2 = 0.50$	Bias (CP95)	0.01 (0.94)	0.02 (0.94)	0.01 (0.95)	0.01 (0.93)
		SSD (ESE)	0.06 (0.06)	0.14 (0.13)	0.06 (0.06)	0.06 (0.06)
	Time (in minutes)		53	46	47	47
	Average number of tests (M1)		5000	1000	2892 (57.8%)	2936 (58.7%)
M2/GP	$\beta_1 = 0.50$	Bias (CP95)	0.01 (0.94)	0.01 (0.95)	0.01 (0.95)	0.01 (0.95)
		SSD (ESE)	0.03 (0.03)	0.06 (0.07)	0.03 (0.03)	0.03 (0.03)
	$\beta_2 = 0.50$	Bias (CP95)	0.01 (0.95)	0.02 (0.93)	0.01 (0.93)	0.01 (0.95)
		SSD (ESE)	0.06 (0.06)	0.14 (0.13)	0.06 (0.06)	0.06 (0.06)
	Time (in minutes)		290	310	315	277
M2/GPP	$\beta_1 = 0.50$	Bias (CP95)	0.01 (0.95)	0.01 (0.95)	0.01 (0.95)	0.01 (0.95)
		SSD (ESE)	0.03 (0.03)	0.06 (0.07)	0.03 (0.03)	0.03 (0.03)
	$\beta_2 = 0.50$	Bias (CP95)	0.01 (0.95)	0.01 (0.92)	0.01 (0.94)	0.01 (0.94)
		SSD (ESE)	0.06 (0.06)	0.14 (0.13)	0.06 (0.06)	0.06 (0.06)
	Time (in minutes)		56	49	49	48
	Average number of tests (M2)		5000	1000	2926 (58.5%)	2956 (59.1%)

Table 1 summarizes the results of estimating β_1 and β_2 in both models (M1 and M2) when assay accuracy probabilities $S_{e(m)}$ and $S_{p(m)}$, $m = 1, 2$, are known. In addition to the three group testing protocols (MPT, DT, and AT), we included the corresponding IT results for comparison. The values of "Bias" and "SSD" in Table 1 are the empirical bias and standard deviation of the $B = 500$ posterior mean estimates, and "ESE" is an averaged estimated posterior standard deviation. For all group testing protocols, the bias in the estimates of β_1 and β_2 is close to zero, SSD and ESE are in close agreement, and the empirical coverage probabilities of 95% equal-tail credible intervals are at the nominal level. This is true regardless of whether a GP or GPP prior model is used, although performing this simulation took about 5–7 times longer using GP. Finally, we note the estimates of β_1 and β_2 from MPT are about 2–3 times more variable than those from IT. This is not unexpected because MPT does not resolve positive pools which leads to a loss in information. However, the same estimates under DT and AT possess the same level of precision as the corresponding estimates from IT, despite DT and AT requiring slightly less than 60% the total number of tests on average.

Table 2. *Simulation results for models M1 and M2 under GP and GPP priors when assay accuracy probabilities are unknown. 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. Note that close agreement between SSD and ESE is preferred. The number of individuals is* $N = 5000$.

Model			$\beta_1 = 0.50$	$\beta_2 = 0.50$	$S_{e(1)} = 0.95$	$S_{p(1)} = 0.98$	$S_{e(2)} = 0.98$	$S_{p(2)} = 0.99$
M1/GP	DT	Bias (CP95)	0.02 (0.92)	0.02 (0.93)	−0.05 (0.94)	−0.01 (0.99)	0.00 (0.98)	0.00 (0.94)
		SSD (ESE)	0.03 (0.04)	0.07 (0.06)	0.04 (0.06)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)
	AT	Bias (CP95)	0.01 (0.94)	0.01 (0.95)	0.00 (0.92)	0.00 (0.96)	0.00 (0.94)	0.00 (0.96)
		SSD (ESE)	0.03 (0.03)	0.06 (0.06)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
M1/GPP	DT	Bias (CP95)	0.02 (0.91)	0.02 (0.93)	−0.05 (0.94)	−0.01 (0.99)	0.00 (0.98)	0.00 (0.94)
		SSD (ESE)	0.04 (0.04)	0.07 (0.06)	0.04 (0.06)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)
	AT	Bias (CP95)	0.01 (0.95)	0.01 (0.93)	0.00 (0.93)	0.00 (0.97)	0.00 (0.95)	0.00 (0.97)
		SSD (ESE)	0.03 (0.03)	0.06 (0.06)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
M2/GP	DT	Bias (CP95)	0.02 (0.92)	0.02 (0.95)	−0.05 (0.92)	−0.01 (0.98)	0.00 (0.98)	0.00 (0.96)
		SSD (ESE)	0.04 (0.04)	0.06 (0.06)	0.04 (0.05)	0.01 (0.02)	0.01 (0.01)	0.00 (0.01)
	AT	Bias (CP95)	0.01 (0.95)	0.01 (0.96)	0.00 (0.95)	0.00 (0.96)	0.00 (0.94)	−0.01 (0.95)
		SSD (ESE)	0.03 (0.03)	0.06 (0.06)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
M2/GPP	DT	Bias (CP95)	0.02 (0.92)	0.02 (0.95)	−0.05 (0.93)	−0.01 (0.98)	0.00 (0.98)	0.00 (0.97)
		SSD (ESE)	0.04 (0.04)	0.06 (0.06)	0.04 (0.05)	0.01 (0.02)	0.01 (0.01)	0.00 (0.01)
	AT	Bias (CP95)	0.01 (0.96)	0.01 (0.95)	0.00 (0.95)	0.00 (0.96)	0.00 (0.95)	−0.01 (0.95)
		SSD (ESE)	0.03 (0.03)	0.06 (0.06)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)

Table 2 presents the same summaries as in Table 1 except that now the assay accuracy probabilities $S_{e(1)}$, $S_{p(1)}$, $S_{e(2)}$, and $S_{p(2)}$ are treated as unknown and are estimated along with the regression parameters β_1 and β_2. Therefore, only DT and AT are shown in Table 2 as only these protocols implement both MPT and IT. In terms of estimation and inference, our findings for the regression parameters β_1 and β_2 in this setting are analogous to those in Table 1. For the assay accuracy probabilities, which were modeled *a priori* using uniform distributions, there is evidence the master pool sensitivity $S_{e(1)}$ is slightly underestimated for DT on average. However, this does not occur when AT is used, and inferences for the other accuracy probabilities $S_{p(1)}$, $S_{e(2)}$, and $S_{p(2)}$ are all on target.

The fundamental difference between the methods in this article and other group testing regression approaches is our ability to estimate the unknown functions $g_l(\cdot)$ in (2.1). Figure 1 shows the results when estimating $g_2(\cdot)$ in Model 1 in our simulation study, assuming assay accuracy probabilities are known, by employing both GP and GPP priors. The same figure for $g_1(\cdot)$ in Model 1, the same figures for Model 2 (assuming known accuracies), and the same figures for both Models 1 and 2 (assuming unknown accuracies) are shown in the supplementary material available at *Biostatistics* online. In all figures, we display the 0.025, 0.50, and 0.975 quantiles of the $B = 500$ estimated functions (posterior means) from our simulation. With the exception of MPT, the median estimated functions are in nearly perfect agreement with the true regression functions for both the GP and GPP prior models. Estimates calculated from DT and AT exhibit less variability than those calculated from MPT and appear to be as efficient as those from IT despite requiring far fewer tests.

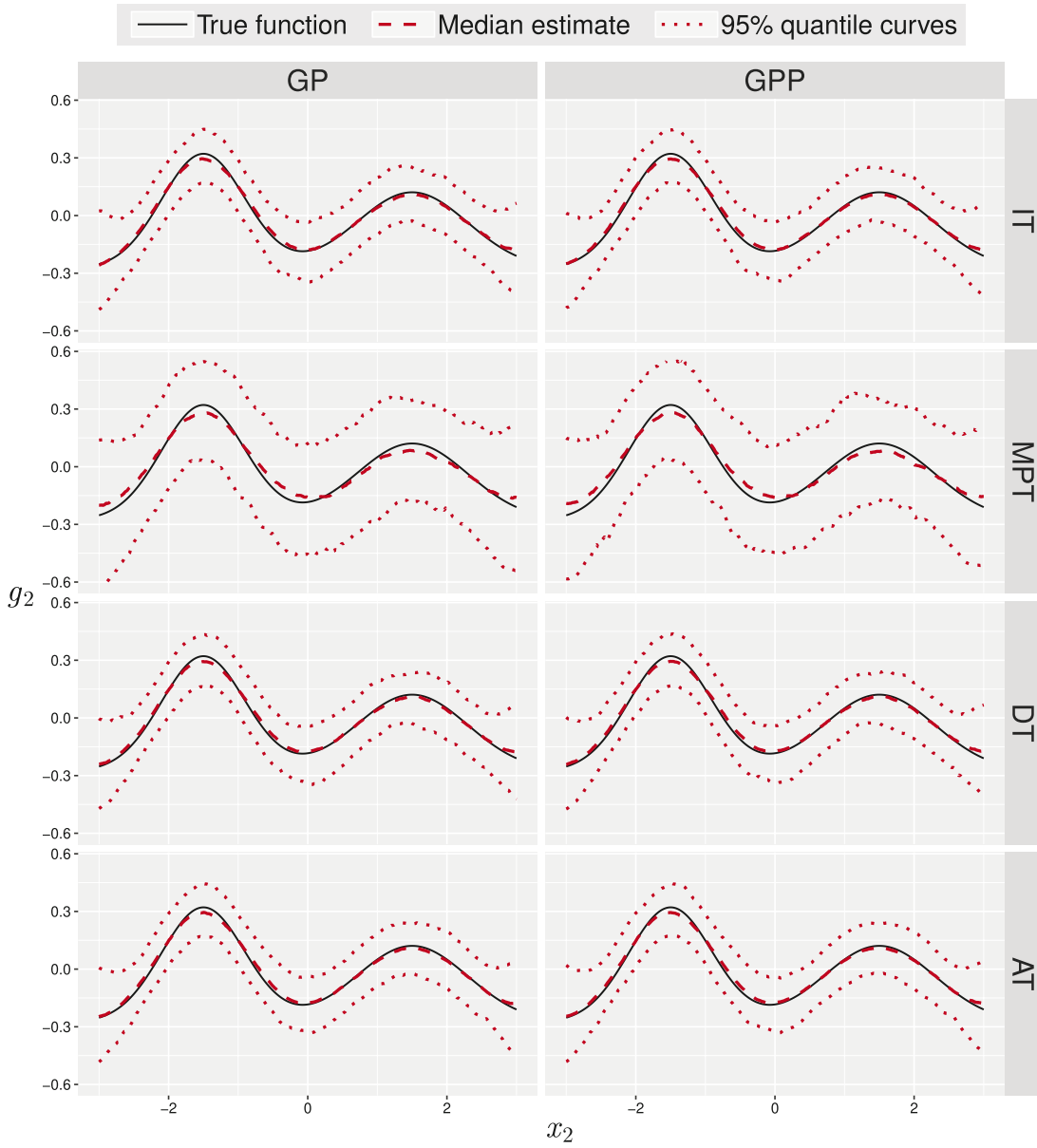

Fig. 1. Simulation results for IT, MPT, DT, and AT when assay accuracy probabilities are known. The solid curve in each subfigure is the second function $g_2(x_2)$ in model M1. Both GP (left) and GPP (right) prior models are used for estimation. The following posterior mean quantiles are shown: 0.025 (dotted curve), 0.50 (dashed curve), and 0.975 (dotted curve).

4. Iowa chlamydia data analysis

We apply our regression methods to a data set provided by our colleagues at the State Hygienic Laboratory (SHL) in Coralville, Iowa. The SHL is the largest public health laboratory in Iowa, and each year the lab tests between 20 and 30 thousand residents for chlamydia. The lab normally receives around 100 specimens

each working day from clinics located throughout the state; these specimens (most of which are swab or urine specimens) are shipped to the SHL where testing is performed using the Aptima Combo 2 Assay (Hologic, San Diego). Because the SHL offers its services without charge to the patient or to the clinic responsible for the specimen, testing costs for the lab are an omnipresent concern—especially because federal funds for chlamydia screening have diminished in recent years. The current protocol at the SHL is to use DT for all swab specimens collected from females, usually in master pools of size four, and to use IT for all other specimens. Swab master pools which test positively are resolved immediately so that final diagnoses can be provided to patients in a timely manner.

Like other public health labs, the primary reason the SHL uses group testing is to save money. In fact, our colleagues have reported that pooling female swab specimens saves the lab approximately $600 000 per year when compared to what it would cost to test these specimens individually. However, because chlamydia infection can be asymptomatic, surveillance is also critical to inform public health efforts in Iowa towards reducing the prevalence—especially for those residents at the highest risk. The methodology in this article can be used to analyze data exhibiting the complicated structure like those collected at the Iowa SHL. For illustration, we focus on the diagnoses of $N = 13\,862$ female subjects tested during 2014. These diagnoses are derived from the test results on 2273 swab master pools of size four, 12 swab master pools of size three, 1 swab master pool of size two, 416 individual swab specimens, and 4316 individual urine specimens. Further testing was performed on individuals in the 2286 $(= 2273 + 12 + 1)$ swab master pools in accordance with Dorfman's protocol when necessary.

After an extensive model-building process using all available risk factors collected on each individual, we selected the following model to describe an individual's chlamydia status:

$$H^{-1}\{\text{pr}(\widetilde{Y} = 1|\mathbf{x})\} = \beta_0 + g_1(x_1) + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 x_5 + \beta_5 x_6, \tag{4.7}$$

where x_1 denotes age (in years), x_2 is a race indicator $(= 1$ if Caucasian; $= 0$ otherwise), $x_3 = 1$ if a new sexual partner was reported within the last 90 days (0, otherwise), $x_4 = 1$ if multiple sexual partners were reported within the last 90 days (0, otherwise), $x_5 = 1$ if there was contact with at least one partner who reported an STD within the last 90 days (0, otherwise), and $x_6 = 1$ if there were symptoms of infection reported; e.g., painful urination/menstruation, etc. The unknown function $g_1(\cdot)$ in (4.7) allows the marginal effect of age on chlamydia status to be nonlinear. We continue to assume $H(\cdot)$ is the probit link.

Acknowledging differences in how the Aptima Combo 2 Assay may perform on swab and urine specimens (Gaydos *and others*, 2003) and also allowing for differences between testing pools and testing individuals, we posited three sensitivity and specificity parameter pairs: $S_{e(1)}$ and $S_{p(1)}$ for swab specimens tested in pools, $S_{e(2)}$ and $S_{p(2)}$ for swab specimens tested individually, and $S_{e(3)}$ and $S_{p(3)}$ for individual urine specimens. In our analysis, these six parameters were assigned uniform priors. We also assigned diffuse $\mathcal{N}(0, 1000)$ priors for the six regression parameters in (4.7) and adopted the Matérn correlation function in (3.6) to estimate $g_1(\cdot)$. Priors elicited for the precision parameter τ_1 and $\boldsymbol{\theta}_1 = (\nu, \phi)'$ were identical to those in Section 3. Finally, in the data set provided to us, each individual's age was measured to the nearest hundredth of a year; this admits $K_1 = 2743$ unique values of x_1 making the GP approach too computationally intense to implement. To circumvent this problem, we estimated (4.7) under a GP prior using ages rounded to the nearest tenth of a year, which reduced the number of unique observations to 430. To estimate the model using a GPP prior, we selected $\widetilde{K}_1 = 100$ knots equally spaced between 5.6 and 70.0, the minimum and maximum ages of females tested in 2014, respectively.

The results of our analysis are shown in Table 3 and Figure 2. For comparison purposes, we also included the corresponding results from estimating the same probit model as a GLM (McMahan *and others*, 2017), that is, by assuming the effect of age is linear. Table 3 provides posterior mean estimates and 95% highest posterior density credible intervals for the regression parameters in (4.7) and the six assay accuracy probabilities described in the last paragraph. For these parameters, one will note the differences

Table 3. Iowa chlamydia data. Results from estimating the model in (4.7) using GP and GPP priors. Posterior mean estimates and 95% highest posterior density credible intervals are provided. The GLM fit from McMahan *and others* (2017) is also shown.

		GP		GPP		GLM	
Parameter	Description	Estimate	95% HPD	Estimate	95% HPD	Estimate	95% HPD
β_1	Race	−0.175	(−0.257, −0.083)	−0.177	(−0.253, −0.083)	−0.178	(−0.260, −0.097)
β_2	New partner	0.149	(0.083, 0.218)	0.151	(0.078, 0.217)	0.146	(0.083, 0.218)
β_3	Multiple partners	0.179	(0.084, 0.272)	0.175	(0.080, 0.270)	0.175	(0.083, 0.266)
β_4	Contact with STD	0.769	(0.652, 0.895)	0.770	(0.647, 0.897)	0.772	(0.634, 0.898)
β_5	Symptoms	0.155	(0.085, 0.234)	0.154	(0.085, 0.233)	0.148	(0.076, 0.235)
$S_{e(1)}$	Swab pool	0.895	(0.770, 1.000)	0.905	(0.782, 1.000)	0.883	(0.739, 1.000)
$S_{e(2)}$	Swab individual	0.998	(0.994, 1.000)	0.998	(0.994, 1.000)	0.998	(0.994, 1.000)
$S_{e(3)}$	Urine individual	0.845	(0.674, 0.999)	0.850	(0.694, 1.000)	0.818	(0.646, 0.999)
$S_{p(1)}$	Swab pool	0.999	(0.998, 1.000)	0.999	(0.998, 1.000)	0.999	(0.998, 1.000)
$S_{p(2)}$	Swab individual	0.977	(0.963, 0.992)	0.976	(0.961, 0.991)	0.979	(0.964, 0.994)
$S_{p(3)}$	Urine individual	0.988	(0.976, 1.000)	0.987	(0.975, 1.000)	0.988	(0.973, 1.000)

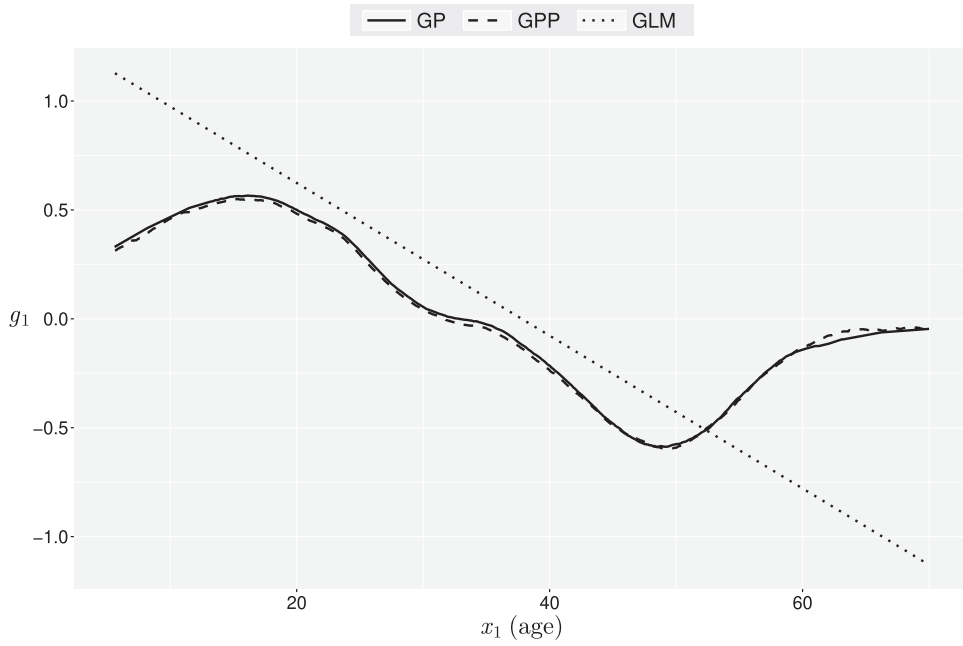

Fig. 2. Iowa chlamydia data. Estimated age effect (i.e., pointwise posterior mean) obtained by GP (solid curve) and GPP (dashed curve) prior models. The GLM fit from McMahan *and others* (2017) is also shown (dotted line).

between the GP and GPP posterior means are small and using a generalized additive model produces estimates which are similar to those when using a GLM. The amount of variability in these estimates is also similar despite the fact that the GP/GPP approaches estimate $g_1(\cdot)$ nonparametrically.

On the other hand, Figure 2 highlights the practical limitations of using group testing regression methods which assume linear covariate effects. Although the effect of age on chlamydia status is approximately linear over a subset of the ages, the age groups where the effect is highly nonlinear correspond to two cohorts of epidemiological importance. First, the Centers for Disease Control and Prevention (CDC) recommends regular chlamydia screening for all females in the United States aged 25 and younger. This is a priority subpopulation for chlamydia prevention given the high burden of risk and the potential for severe complications (e.g., infertility, ectopic pregnancy, etc.). Our analysis of the Iowa data reveals a peak risk of infection around 18 years of age, which is largely consistent with other states (CDC, 2019). Second, a compelling finding from our analysis is the relative increase in chlamydia risk for females aged 50 and older. This observation was initially surprising; however, our public health colleagues at the CDC have noted this group's increase may be part of an emerging national trend involving older adults.

5. Discussion

In this article, we have proposed a generalized additive regression framework for potentially misclassified group testing data, extending the methodology in McMahan *and others* (2017) to incorporate nonlinear covariate effects. GP and GPP prior distributions can be used to estimate the unknown functions describing these effects, and our careful use of data augmentation leads to a computationally efficient posterior sampling algorithm. Simulation results consistently demonstrate that our approach can reliably estimate the regression parameters and unknown functions even when assay classification accuracy probabilities

are unknown. Our methods are applied to group testing data collected in Iowa, where pooling is used to reduce the cost of testing individuals for chlamydia infection.

As noted by an anonymous referee, it is important to emphasize the data we analyzed in Section 4 are not data from a random sample of all Iowa females. In fact, the lead lab technician at the SHL has described how specimens received each day are likely representative of the "highest-risk" residents in the state. Of course, this observation does not intrinsically void the value of our regression methodology, but it does limit our ability to make statements about females in lower-risk groups. One possible limitation of our approach is the requirement that testing results are conditionally independent given the true disease statuses; see Section 2.2. Such a requirement is needed whenever individuals appear in multiple pools, as is the case with the Iowa data analysis. In the supplementary materials available at *Biostatistics* online, we have performed an additional simulation study to assess the robustness of our methods to violations of this assumption. Even though this study suggests estimation is largely unaffected by such a violation, finding a way to relax this assumption could be a worthwhile topic for future research.

Finally, it should be possible to develop other group testing regression methods using the latent data framework presented in this article. One possible extension could involve accounting for spatial or spatiotemporal dependence when individuals are tested in pools. This could be accomplished by adopting GP models for point process data (Banerjee *and others*, 2008) or conditional autoregressive models for areal data (Banerjee *and others*, 2014). However, in either case, we would expect the computational difficulty to far exceed that which is seen here. Another useful extension would be to develop joint modeling methods which incorporate testing responses from multiplex assays; i.e., assays which provide responses for multiple diseases at once. We have found that public health labs are increasingly relying on multiplex assays to save resources (when compared to using separate assays for each disease) and that often these assays are applied to pools of individual specimens.

6. Software

Software in the form of R code, a simulated data set, and documentation are available on GitHub (https://github.com/yanliu5/gam). The simulated data set has the same structure as the DT data set in Section 4. The analysis of the simulated data set is shown in the supplementary material available at *Biostatistics* online.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

We are grateful to the Associate Editor and two referees for their helpful comments on an earlier version of this article. We also thank Jeffrey Benfer and Kristofer Eveland at the State Hygienic Laboratory (University of Iowa) and Dr. Elizabeth Torrone at the Centers for Disease Control and Prevention.
Conflict of Interest: None declared.

Funding

The National Institutes of Health (Grant R01 AI121351); National Science Foundation (Grant OIA-1826715) and the Office of Naval Research (Grant N00014-19-1-2295) to C.S.M.

REFERENCES

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

BANERJEE, S., CARLIN, B. AND GELFAND, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall/CRC.

BANERJEE, S., GELFAND, A., FINLEY, A. AND SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B* **70**, 825–848.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2019). *Sexually Transmitted Disease Surveillance 2018*. Atlanta: U.S. Department of Health and Human Services. www.cdc.gov (doi: 10.15620/cdc.79370) (last accessed January 2, 2020).

CHEN, P., TEBBS, J. AND BILDER, C. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–1278.

CHOUDHURI, N., GHOSAL, S. AND ROY, A. (2007). Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology* **4**, 227–243.

DELAIGLE, A. AND HALL, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika* **102**, 871–887.

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.

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.

DORFMAN, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436–440.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2001). *The Elements of Statistical Learning.* Berlin: Springer.

GAYDOS, C., QUINN, T., WILLIS, D., WEISSFELD, A., HOOK, E., MARTIN, D., FERRERO, D. AND SCHACHTER, J. (2003). Performance of the APTIMA Combo 2 Assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in female urine and endocervical swab specimens. *Journal of Clinical Microbiology* **41**, 304–309.

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.

HUANG, X. AND TEBBS, J. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics* **65**, 710–718.

KIM, H., HUDGENS, M., DREYFUSS, J., WESTREICH, D. AND PILCHER, C. (2007). Comparison of group testing algorithms for case identification in the presence of testing error. *Biometrics* **63**, 1152–1163.

KRAJDEN, M., COOK, D., MAK, A., CHU, K., CHAHIL, N., STEINBERG, M., REKART, M. AND GILBERT, M. (2014). Pooled nucleic acid testing increases the diagnostic yield of acute HIV infections in a high-risk population compared to 3rd and 4th generation HIV enzyme immunoassays. *Journal of Clinical Virology* **61**, 132–137.

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

LIU, A., LIU, C., ZHANG, Z. AND ALBERT, P. (2012). Optimality of group testing in the presence of misclassification. *Biometrika* **99**, 245–251.

MCMAHAN, C., TEBBS, J. AND BILDER, C. (2012). Two-dimensional informative array testing. *Biometrics* **68**, 793–804.

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

Polson, N., Scott, J. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Saá, P., Proctor, M., Foster, G., Krysztof, D., Winton, C., Linnen, J., Gao, K., Brodsky, J., Limberger, R., Dodd, R. and Stramer, S. (2018). Investigational testing for Zika virus among US blood donors. *New England Journal of Medicine* **378**, 1778–1788.

Shi, M., Umbach, D. and Weinberg, C. (2014). Disentangling pooled triad genotypes for association studies. *Annals of Human Genetics* **78**, 345–356.

Van, T., Miller, J., Warshauer, D., Reisdorf, E., Jerrigan, D., Humes, R. and Shult, P. (2012). Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology* **50**, 891–896.

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.

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

Westreich, D., Hudgens, M., Fiscus, S. and Pilcher, C. (2008). Optimizing screening for acute human immunodeficiency virus infection with pooled nucleic acid amplification tests. *Journal of Clinical Microbiology* **46**, 1785–1792.

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

Zhang, B., Bilder, C. and Tebbs, J. (2013). Group testing regression model estimation when case identification is a goal. *Biometrical Journal* **55**, 173–189.

[*Received August 10, 2019; revised January 4, 2020; accepted for publication January 13, 2020*]