Biometrics 000, 000–000 DOI: 000
000 0000

Bayesian Additive Regression Trees for Group Testing Data

Madeleine E. St. Ville[1], Christopher S. McMahan[1*], Joe D. Bible[1], Joshua M. Tebbs[2], and Christopher R. Bilder[3]

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

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

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

email: mcmaha2@clemson.edu

Summary: When screening for low-prevalence diseases, pooling specimens (e.g., urine, blood, swabs, etc.) through group testing has the potential to substantially reduce costs when compared to testing specimens individually. A common goal in pooled testing applications is to estimate the relationship between an individual's true disease status and their individual-level covariate information. However, estimating such a relationship is a nontrivial problem in group testing because these true individual disease statuses are unknown due to the group testing protocol and the potential imperfect testing. While several regression methods have been developed in recent years to accommodate the complexity of group testing data, the functional form of covariate effects is assumed to be known. To avoid model misspecification and biased inference, and to provide a more flexible framework, we propose a Bayesian additive regression trees (BART) 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 unknown functions of covariates and assay classification accuracy probabilities.

Key words: Keywords here

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

1. Introduction

In the field of disease screening, group testing has become a popular alternative to individual testing due to its cost-effectiveness and efficiency. Fundamentally, group testing combines individual specimens (e.g., blood, urine, swab, etc.) to form a pooled specimen that is tested for the presence of disease. In many group testing protocols, individuals contributing to a pooled specimen that tests negatively are classified as negative at the expense of just a single diagnostic assay; and in contrast, positive pools are resolved through further testing. It has become a mainstreem approach to screen for a variety of infectious diseases such as HIV (Westreich et al., 2008; Krajden et al., 2014), gonorrhea and chlamydia (Lewis et al., 2012), influenza (Van et al., 2012), Zika (Saá et al., 2018), tuberculosis (Abdurrahman et al., 2015), and SARS-CoV-2 (Torres et al., 2020; Abdulhamid et al., 2020).

Prevalence estimation for disease surveillance involves predicting the probability of disease for individuals and identifying associated risk factors. However, accurately estimating the relationship between an individual's disease status and their covariate information from group testing data is a nontrivial problem; the true individual responses are obscured by the group testing protocol, and their testing responses are potentially misclassified due to imperfect testing. Because of group testing's ability to provide substantial cost and time savings, there has been a growth in the development of regression methods for group testing data to address these challenges. Prominent research includes parametric approaches by Vansteelandt et al. (2000), Huang and Tebbs (2009), Chen et al. (2009), and Delaigle and Tan (2023) as well as semiparametric and nonparametric approaches by Delaigle and Meister (2011), Delaigle et al. (2014), and Delaigle and Hall (2015). More recently, McMahan et al. (2017) proposed a Bayesian approach within a generalized linear model (GLM) framework that boasts three strengths; namely it can analyze data arising from any group testing protocol to include retesting information, it can incorporate historical information about disease prevalence,

and it allows for the estimation of assay accuracy probabilities. McMahan et al. (2017) ultimately motivated the development of several modeling extensions (Joyner et al., 2020; Liu et al., 2021). However, the main limitation of these existing methods is that they require the explicit specification of the functional form of the relationship between covariates and disease status. This can be a challenging task, especially when the relationship is complex. Nonlinear or high-order interaction effects are potentially ignored, which can result in model misspecification and biased inference.

In this article, we propose a Bayesian additive regression trees (BART) modeling framework to estimate regression models using group testing data. BART is an ensemble, machine learning technique that employs a tree-based, nonparametric approach and is well-equipped to handle large, complex group testing data sets. It builds a series of decision trees that partition the input space into regions and makes predictions based on the covariates values within each region. It automatically captures any complex, high-order interaction effects without requiring the researcher to specify anything about the functional form. Furthermore, BART extends the ensemble of decision trees by incorporating Bayesian modeling in order to quantify uncertainty in parameter estimates and regularize the fit. Our proposed BART approach addresses the limitations of existing group testing regression methods, while maintaining the strengths of McMahan et al. (2017), to allow for a more flexible, robust modeling technique that yields more accurate predictions and a better understanding of the relationship between covariates and disease status.

The remainder of this article is organized as follows. In Section 2, we introduce the proposed BART model and describe modeling assumptions. In Section 3, we describe the data augmentation steps that facilitate our Bayesian framework and introduce our posterior sampling algorithm. In Section 4, we present the results of multiple simulations to assess the performance of our proposed method under a variety of settings for group testing protocols.

In Section 5, we present analysis results for chlamydia testing data to illustrate the proposed technique. Finally, in Section 6, we conclude with a summary discussion and describe future research.

2. Notation and Model Formulation

Consider a setting in which a group testing protocol is used to screen N individuals for a binary characteristic of interest, such as the presence or absence of a disease. Let $\widetilde{Y}_i$, for $i = 1, ..., N$, denote the true disease status of the ith individual, with the usual convention that $\widetilde{Y}_i = 1$ indicates that the individual is truly positive, and $\widetilde{Y}_i = 0$ otherwise. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iQ})'$ denote a vector of Q covariates observed for the ith individual. For ease of exposition, we aggregate the individuals' true infection statuses into the vector $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_N)'$, and their covariates into the matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$. For modeling purposes, we assume that the individuals' true disease statuses are conditionally independent given their individual-level covariate information, and assume that the relationship between $\widetilde{Y}_i$ and $\boldsymbol{x}_i$ is given by

$$\Phi^{-1}\left\{P(\widetilde{Y}_i = 1 \big| \boldsymbol{x}_i)\right\} = f(\boldsymbol{x}_i), \tag{1}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution (i.e., the probit link function), $f(\cdot)$ is an unknown function and represents an infinite-dimensional parameter. To reduce its dimension while also maintaining adequate modeling flexibility, we will approximate $f(\cdot)$ using Bayesian additive regression trees (BART); i.e., we approximate $f(\cdot)$ by an ensemble of K regression trees in the following manner:

$$f(\boldsymbol{x}_i) \approx \eta(\boldsymbol{x}_i) := \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k, M_k), \tag{2}$$

where T_k is the kth regression tree structure consisting of a set of interior node decision rules and a set of b_k terminal nodes; $M_k = (\mu_{k1}, \ldots, \mu_{kb_k})'$ is a b_k-dimensional vector of terminal node parameters; and $g(\boldsymbol{x}_i; T_k, M_k)$ is a function that returns the value $\mu_{kt} \in M_k$ if

$\boldsymbol{x}_i$ is assigned to the t-th terminal node based on the interior node decision rules of T_k. The decision rules provide information on which covariate to split on and the associated cutoff value. They are binary splits based on a single predictor, and are of the form $\{x_{iq} \leqslant c\}$ versus $\{x_{iq} > c\}$ for some cutoff values c. Taken together, $g(\boldsymbol{x}_i; T_k, M_k)$ can be viewed as a multi-dimensional step function that can aptly account for many features; e.g., nonlinear effects and interactions of varying orders. Note that in model (2), K is the (typically fixed) number of regression trees. Setting K to be large is recommended for flexible estimation and, through a wide variety of simulated examples, Chipman et al. (2010) showed that the default K=200 yields good predictive performance.

To illustrate the main idea of a sum-of-trees model, consider the following example of an ensemble of K=2 trees and Q=3 covariates. Suppose we are given the two trees in Figure 1. Each tree uses two predictors to split the data into subgroups; the first tree ($k = 1$) on the left of Figure 1 uses x_{i1} and x_{i2}, while the second tree ($k = 2$) on the right uses x_{i3} and x_{i2}. For each tree and for each individual, every value of $\boldsymbol{x}_i$ is assigned to a single terminal node by following a sequence of decision rules at each interior node from top to bottom where it is finally assigned a parameter value associated with that terminal node. Consider the hypothetical data from 5 subjects given in Table 1. We can see that the quantity being 'summed' in the final sum-of-trees model for the ith subject is the terminal node parameter value that each tree structure assigns to this ith subject.

[Figure 1 about here.]

[Table 1 about here.]

BART casts the sum-of-trees model into the Bayesian paradigm and controls the size and effect of the individual trees by imposing regularization priors (Chipman et al., 2010). If the individuals' true disease statuses were observed, we could fit the BART model via standard statistical software; e.g., `Bayestree` (Chipman et al., 2010), `bartMachine` (Kapelner and

Bleich, 2016), and `bart` (Sparapani et al., 2021). However, due to the effects of both pooling and imperfect testing, the individuals' true infection statuses are unobservable in the group testing setting. The observed data available for model fitting consist of error-contaminated test responses that are taken on pools and/or individuals according to a group testing protocol. Further complicating the data structure, many group testing protocols require individuals to be tested in multiple, possibly overlapping, pools (Gastwirth and Johnson, 1994; Johnson and Gastwirth, 2000; Krajden et al., 2014).

To maintain generality and accommodate data from any group testing protocol, we track pool membership through the index sets $\mathcal{P}_j \subset \{1, 2, \ldots, N\}$, where $\mathcal{P}_j$ consists of the indices of individuals who contributed to the jth pool, $j = 1, ..., J$. Let Z_j denote the test outcome observed from assaying the jth pool, with the convention that $Z_j = 1$ denotes the event that the pool tested positively, and $Z_j = 0$ otherwise. To relate the test outcomes to the individual-level covariates, we assume that $S_{ej} = P(Z_j = 1 \mid \widetilde{Z}_j = 1)$ and $S_{pj} = P(Z_j = 0 \mid \widetilde{Z}_j = 0)$, where S_{ej} and S_{pj} are the sensitivity and specificity of the assay when used to test the jth pool, and where $\widetilde{Z}_j$ is the true status of the jth pool. A few comments are warranted. First, the true status of a pool is said to be positive ($\widetilde{Z}_j = 1$) if it contains at least one truly positive individual, and negative ($\widetilde{Z}_j = 0$) otherwise; i.e., $\widetilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0)$. Second, we consider pool-specific assay accuracies (i.e., S_{ej} and S_{pj}) to account for changes in these measures that are related to the use of different assays or other factors that could impact the assay's performance; e.g., specimen type, pool size (i.e., cardinality of $\mathcal{P}_j$).

In some settings, it may be reasonable to assume that the assay accuracies (i.e., S_{ej} and S_{pj}) are known *a priori*. In other settings, however, this may be an untenable assumption. In the latter case, we can estimate these parameters following the approach of McMahan et al. (2017). To do so, we first divide the test outcomes into L different strata based on relevant factors; e.g., pool size and specimen/assay type. With this, define the index set $\mathcal{M}(l) = \{j :$

the jth test outcome is a part of the lth strata}. We assume that the test accuracies vary across these L strata, but are constant within strata. Thus, define $S_{e(l)}$ and $S_{p(l)}$ to be the sensitivity and specificity of the assay associated with the lth strata; i.e., $S_{ej} = S_{e(l)}$ and $S_{pj} = S_{p(l)}$ if and only if $j \in \mathcal{M}(l)$. Proceeding in this fashion leads to a straightforward way of estimating these unknown quantities as well as a way to inject information about them through prior specifications; for further discussion, see Sections 4 and 5.

Based on the relations outlined above, the conditional distribution of $\boldsymbol{Z} = (Z_1, ..., Z_J)'$ is given by

$$\pi(\boldsymbol{Z}\big|\boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = \sum_{\widetilde{\boldsymbol{Y}} \in \{0,1\}^N} \left[\prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \times \prod_{i=1}^{N} \{\Phi(\eta_i)\}^{\widetilde{Y}_i} \{1 - \Phi(\eta_i)\}^{1-\widetilde{Y}_i} \right], \tag{3}$$

where $\boldsymbol{S_e} = (S_{e(1)}, ..., S_{e(L)})'$, $\boldsymbol{S_p} = (S_{p(1)}, ..., S_{p(L)})'$, $\boldsymbol{T} = (T_1, ..., T_K)'$, $\boldsymbol{M} = (M_1, \ldots, M_K)'$, and $\eta_i = \eta(\boldsymbol{x}_i)$. In order to derive (3), we assume that the observed testing responses $\boldsymbol{Z}$ are conditionally independent given their true statuses $\widetilde{\boldsymbol{Z}} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_J)'$, and that $\boldsymbol{Z} \mid \widetilde{\boldsymbol{Z}}$ does not depend on the covariates $\boldsymbol{X}$. These assumptions are common among the group testing literature; e.g., see Vansteelandt et al. (2000); Xie (2001). Note that evaluating the data model outlined in (3) requires taking the sum over the set $\{0, 1\}^N$, which denotes the collection of all 2^N possible realizations of $\widetilde{\boldsymbol{Y}}$. For this reason, directly evaluating (3) can be computationally burdensome, if at all feasible. Admittedly, under specific group testing strategies (e.g., master pool testing) simplifications are possible, yet not in the general case. To overcome this limitation, we make use of a data augmentation strategy, described in Section 3.1, to develop a posterior sampling algorithm that circumvents the need to directly evaluate this data model.

2.1 *Prior specifications*

To complete our Bayesian model, we specify priors for each of the unknown model parameters; i.e., the parameters governing the sum-of-trees model and the testing assay accuracies. Recall, the sum-of-trees model (2) is determined by K trees $(T_1, M_1), \ldots, (T_K, M_K)$. Thus, we must impose priors on the kth tree structure, T_k, and the terminal node parameters given this kth tree structure, $M_k \mid T_k$, for each $k = 1, \ldots, K$. Assuming that the trees $(T_1, M_1), \ldots, (T_K, M_K)$ are independent of each other, we can write the prior distribution as

$$\begin{aligned} \pi\left\{(T_1, M_1), \ldots, (T_K, M_K)\right\} &= \prod_{k=1}^{K} \pi(T_k, M_k) = \prod_{k=1}^{K} \pi(M_k \mid T_k)\pi(T_k) \\ &= \prod_{k=1}^{K} \prod_{t=1}^{b_k} \pi(\mu_{kt} \mid T_k)\pi(T_k), \end{aligned} \tag{4}$$

noting that the last line of (4) follows from the assumption that the terminal node parameters are conditionally independent given their tree structure. To elicit priors for each T_k and $\mu_{kt} \mid T_k$, we will follow the work of Chipman et al. (2010), which we briefly outline below. These prior specifications are simplified by using identical forms for all $\pi(T_k)$ and for all $\pi(\mu_{kt} \mid T_k)$, for $t = 1, \ldots, b_k$, and for $k = 1, \ldots, K$.

We first specify $\pi(T_k)$, the prior on the kth tree structure, based on three probabilistic rules that control the size (i.e., number of terminal nodes) of the tree, the variables to split on, and the locations of the split. The size of the tree is based on the depth of the terminal nodes, where a node at depth $d \in \{0, 1, 2, \ldots\}$ is nonterminal (i.e., an interior node) with probability $\alpha(1 + d)^{-\beta}$, where $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$. The default values of the hyperparameters recommended by Chipman et al. (2010), and used herein, are $\alpha = 0.95$ and $\beta = 2$. These default specifications tend to *a priori* favor smaller trees; i.e., trees having 2 to 3 terminal nodes. For nonterminal nodes, the variable to split on is randomly selected from the set of available covariates, and the location of the split, given the selected variable, is again sampled at random from the set of observed values of the selected variable. Next, the prior specification for the terminal node parameters is given as $\mu_{kt} \sim N(0, \sigma_\mu^2)$, where

$\sigma_\mu = 3.0/\left(H\sqrt{K}\right)$, and $H = 2$ is the recommended default hyperparameter value, which we will adopt herein. The aim of this prior is to provide model regularization: it has the ability to shrink the terminal node parameters, limiting the effect of the individual tree components. For further details, see Chipman et al. (2010).

Finally, to acknowledge uncertainty in the assay accuracies, we must also elicit prior distributions for $S_{e(l)}$ and $S_{p(l)}$, $l = 1, \ldots, L$. Given the form of the data model (3), we naturally specify the following independent Beta priors:

$$\begin{aligned} S_{e(l)} &\sim Beta(a_{e(l)}, b_{e(l)}) \\ S_{p(l)} &\sim Beta(a_{p(l)}, b_{p(l)}), \text{ for } l = 1, \ldots, L. \end{aligned} \tag{5}$$

When historical information about assay performance is available (e.g., from pilot studies used to validate the testing assay), we can incorporate it into the model by choosing hyperparameter values that reflects this prior knowledge. We illustrate this strategy in Section 5.

3. Posterior Inference

3.1 *Data augmentation*

Recall that evaluating the data model (3) is computationally infeasible. To facilitate the development of an efficient posterior sampling algorithm and to avoid direct evaluation of (3), we propose a two-stage data augmentation procedure. In the first stage, we introduce the individuals' true disease statuses $\widetilde{Y}_i$ as latent random variables. This leads to the following joint conditional distribution:

$$\begin{aligned} \pi(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}} | \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = &\prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \\ &\times \prod_{i=1}^{N} \{\Phi(\eta_i)\}^{\widetilde{Y}_i} \{1 - \Phi(\eta_i)\}^{1-\widetilde{Y}_i}. \end{aligned} \tag{6}$$

Making use of the fact that our model employs the probit link function $\Phi(\cdot)$, the second stage of our data augmentation strategy introduces a carefully constructed latent random

variable, ω_i, for each individual, for $i = 1, \ldots, N$. These random variables independently obey a standard normal distribution such that $\omega_i > 0$ if $\widetilde{Y}_i = 1$ and $\omega_i \leqslant 0$ if $\widetilde{Y}_i = 0$; for further details see Albert and Chib (1993). This stage of our data augmentation procedure yields the following augmented likelihood:

$$\pi(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}}, \boldsymbol{\omega} \big| \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{M}) = \prod_{l=1}^{L} \prod_{j \in \mathcal{M}(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{\widetilde{Z}_j} \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{1-\widetilde{Z}_j} \times \prod_{i=1}^{N} \phi(\omega_i - \eta_i) \left\{ I(\widetilde{Y}_i = 1, \omega_i > 0) + I(\widetilde{Y}_i = 0, \omega_i \leqslant 0) \right\}, \quad (7)$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)'$, and $\phi(\cdot)$ denotes the standard normal probability density function. This two-stage data augmentation procedure, together with the proposed prior specifications, allows for the construction of a full Gibbs sampling algorithm to be used for posterior inference.

3.2 *Posterior sampling algorithm*

In this section, we briefly describe the full conditional distributions used in our posterior sampling algorithm. A complete, step-by-step description of the posterior sampling algorithm is provided in Web Appendix C.

Attention is first turned to the latent random variables introduced through the data augmentation procedure. It follows from (7) that the full conditional of $\widetilde{Y}_i$ is given as $\widetilde{Y}_i \big| \boldsymbol{Z}, \widetilde{\boldsymbol{Y}}_{-i}, \boldsymbol{S_e}, \boldsymbol{S_p}, \boldsymbol{T}, \boldsymbol{M} \sim \textit{Bernoulli}\left\{p_{i1}^* / (p_{i0}^* + p_{i1}^*)\right\}$, where $\widetilde{\boldsymbol{Y}}_{-i}$ is the vector $\widetilde{\boldsymbol{Y}}$ with the ith element removed, and

$$p_{i1}^* = \Phi(\eta_i) \prod_{l=1}^{L} \prod_{j \in \mathcal{I}_i(l)} S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j}$$

$$p_{i0}^* = \{1 - \Phi(\eta_i)\} \prod_{l=1}^{L} \prod_{j \in \mathcal{I}_i(l)} \left\{ S_{e(l)}^{Z_j} (1 - S_{e(l)})^{1-Z_j} \right\}^{I(s_{ij} > 0)} \left\{ (1 - S_{p(l)})^{Z_j} S_{p(l)}^{1-Z_j} \right\}^{I(s_{ij} = 0)},$$

where $s_{ij} = \sum_{i' \in \mathcal{P}_j : i' \neq i} \widetilde{Y}_{i'}$ and $\mathcal{I}_i(l) = \{j \in \mathcal{M}(l) : i \in \mathcal{P}_j\}$. Further, it follows from (7) that the full conditional of ω_i is truncated normal, where the truncation depends on the ith latent

disease status $\widetilde{Y}_i$; that is,

$$\omega_i \mid \widetilde{Y}_i, \boldsymbol{T}, \boldsymbol{M} \sim \begin{cases} TN\{\eta_i, 1, (0, \infty)\}, & \text{if } \widetilde{Y}_i = 1 \\ TN\{\eta_i, 1, (-\infty, 0)\}, & \text{if } \widetilde{Y}_i = 0, \end{cases} \tag{8}$$

for $i = 1, \ldots, N$, where $TN\{\mu, \sigma^2, (a, b)\}$ denotes a truncated normal distribution with mean μ, variance σ^2, and support over the interval (a, b).

Given the carefully constructed latent variables and the form of the augmented likelihood (7), sampling the sum-of-trees model parameters and the assay accuracies is straightforward. In particular, we follow the Bayesian backfitting algorithm of Chipman et al. (2010) to sample all parameters associated with the K regression trees. Refer to Web Appendices A and B for details and complete expressions of the posteriors for the sum-of-trees model parameters. Furthermore, given the Beta prior specifications (5) imposed on the assay accuracies, their full conditional distributions are also Beta; i.e.,

$$S_{e(l)} \mid \boldsymbol{Z}, \widetilde{\boldsymbol{Y}} \sim Beta(a^*_{e(l)}, b^*_{e(l)})$$

$$S_{p(l)} \mid \boldsymbol{Z}, \widetilde{\boldsymbol{Y}} * \sim Beta(a^*_{p(l)}, b^*_{p(l)}), \text{ for } l = 1, \ldots, L,$$

where $a^*_{e(l)} = a_{e(l)} + \sum_{j \in \mathcal{M}(l)} Z_j \widetilde{Z}_j$, $b^*_{e(l)} = b_{e(l)} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j) \widetilde{Z}_j$, $a^*_{p(l)} = a_{p(l)} + \sum_{j \in \mathcal{M}(l)} (1 - Z_j)(1 - \widetilde{Z}_j)$, and $b^*_{p(l)} = b_{p(l)} + \sum_{j \in \mathcal{M}(m)} Z_j (1 - \widetilde{Z}_j)$.

3.3 *Variable Selection*

After a suitable burn-in period, our posterior algorithm returns S posterior samples of the K regression tree structures. A byproduct of our BART approach is its ability to provide a measure of variable importance, following the approach of Chipman et al. (2010). Within a posterior sample, we can compute the proportion of times a particular covariate is used as a splitting variable among all decision rules in the ensemble of K trees. We can then estimate the variable inclusion proportion for this covariate as the posterior mean of these proportions across the S posterior samples.

In particular, let z_{qs} be the number of decision rules that use the qth covariate as the splitting variable in the sth posterior draw of the sum-of-trees model, and let $z_{\cdot s} = \sum_{q=1}^{Q} z_{qs}$ be the total number of decision rules in this sth posterior sample. With this, we define

$$v_q = \frac{1}{S} \sum_{s=1}^{S} \frac{z_{qs}}{z_{\cdot s}} \tag{9}$$

to be the variable inclusion proportion for the qth covariate. Intuitively, covariates with large variable inclusion proportions are identified as the more influential predictors of the outcome of interest. Therefore, through the use of these inclusion proportions, covariates can be ranked in terms of their relative importance in the prediction of the outcome. This strategy is more effective when the number of trees K is small because predictors will be forced to compete with each other to improve the fit (Chipman et al., 2010). We illustrate this variable importance strategy in Sections 4 and 5.

This approach is widely used throughout BART literature, but it has its limitations as a result of BART's tendency to overfit noise (Bleich et al., 2014). Hence, it is not appropriate to use these inclusion proportions as a way to select a subset of important predictors. There are many other approaches that improve upon this strategy, but they are beyond the scope of this work.

4. Simulation Studies

To evaluate the performance of our proposed BART technique, we will conduct numerical studies with simulated data designed to mimic the primary features of the Iowa chlamydia data application discussed in Section 5. We consider two population-level models, both of which follow the form of (1), where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})'$ is a vector of Q=3 covariates with $x_{i1}, x_{i2} \overset{\text{ind.}}{\sim} \text{Uniform}(0, 10)$, and $x_{i3} \sim \textit{Bernoulli}(0.5)$. The first model (M1) is defined as

$$f(\boldsymbol{x}_i) = \sin(\pi \cdot x_{i1}) - 1.25,$$

while the second model (M2) is

$$f(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' = (-0.85, 0.55, -1.25, -0.35)'$. These population models induce population prevalences that are consistent with the chlamydia prevalence observed in the motivating data. The first model M1 was chosen to evaluate BART's performance when the data exhibits nonlinear patterns, and the second model M2 was chosen to assess potential losses from a BART fit when a conventional linear model would be appropriate. For each model M1 and M2, we generated N=5000 individual true statuses as $\widetilde{Y}_i \sim Bernoulli(\Phi(f(\boldsymbol{x}_i))$, where $f(\boldsymbol{x}_i)$ took the form under each respective population model. This sample size was chosen to be roughly on third of the motivating data sample size. We used this process to simulate 500 independent, individual-level datasets.

We generated the pooled testing outcomes under two group testing protocols: master pool testing (MPT) and Dorfman testing (DT). With MPT, individuals are assigned to exactly one master pool for testing but no further testing is performed, and hence is for estimation purposes only. DT is a two-stage hierarchical procedure that completes the testing process of MPT by individually retesting all members contributing to positive master pools. For both of these protocols, the individual true statuses $\widetilde{Y}_i$ were randomly assigned to master pools of size 4, following the pooling procedure used in the motivating data, and the testing response for the jth pool was generated as $Z_j \mid \widetilde{Z}_j \sim Bernoulli\{S_{ej}\widetilde{Z}_j + (1 - S_{pj})\}$, where $\widetilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0)$. For the assay accuracies, two different simulation configurations were considered. In the first setting, $S_{ej} = 0.95$ and $S_{pj} = 0.98$ for all $j = 1, \ldots, J$, and we assume sensitivity and specificity are known *a priori*. In the second setting, only DT is implemented and assay accuracy varies across L=2 testing outcome strata: outcomes test in pools ($l = 1$) and outcomes individually tested ($l = 2$). Specifically, $S_{e(1)} = 0.95$, $S_{p(1)} = 0.98$,

and $S_{e(2)} = 0.98$, $S_{p(2)} = 0.99$. Under this setting, the assay accuracies are considered to be unknown and estimated.

Our proposed approach is evaluated under two BART configurations: one with K=20 trees and another with K=200 trees. For the model parameters associated with the K regression trees, we used the default prior specifications described in Section 2.1. Under the unknown assay accuracy setting, we assumed that no prior knowledge about assay testing performance is available and elicited flat, uninformative Beta priors for the accuracy probabilities; i.e., $S_{e(l)}, S_{p(l)} \sim Beta(1, 1)$. We used our posterior sampling algorithm to draw 2500 posterior samples after a burn-in of 2500 samples. Convergence of the chains was assessed using standard MCMC diagnostics. As a competitive technique, we also fit the Bayesian generalized linear model (GLM) from McMahan et al. (2017) using all three covariates for all datasets, where flat priors were specified for all regression coefficients. This GLM fit is incorrectly specified for M1 and correctly specified for M2. Proceeding in this fashion allows us to examine the pros and cons of BART, both when reasonable and when GLM is more appropriate.

All estimation results were averaged over the 500 independent datasets. To evaluate in- and out-of-sample classification accuracy, we conducted a receiver operating characteristic (ROC) curve analysis, summarized by the area under the curve (AUC). To assess out-of-sample classification accuracy for each model fit, we simulated 1,000 new individuals using the process outlined above and then used our models to predict their infection probabilities and compute the associated AUC scores. To illustrate BART's variable selection strategy, the variable inclusion proportions (9), were recorded for each covariate based on the S=2500 MCMC samples. For purposes of comparison, individual testing (IT) was also implemented for the simulation configurations which assume that the common assay accuracies $S_{ej} = 0.95$ and $S_{pj} = 0.98$ are known.

Figure 2 shows the in-sample data results from estimating $f(\cdot)$ in M1 when the assay accuracies are known. The estimated functions from both BART configurations are in agreement with the true regression function of population model M1 and there is not a significant difference in the estimated functions between the two configurations, suggesting that BART is more or less robust to the number of trees. There is a massive improvement in the estimated functions from the DT protocol compared to MPT, which is not surprising because of the loss of information in MPT. Overall, Figure 2 showcases BART's ability to model the nonlinear relationship between response and predictor variables, without having to specify what the surface looks like. Unsurprisingly, without any explicit specification of the functional form, the standard GLM falls apart for this nonlinear model.

Web Table 1 in Web Appendix D summarizes the ROC analysis for in- and out-of-sample predictive accuracy under both population models. For M1, BART has substantially better classification accuracy than the incorrectly specified GLM; and for M2, BART performs just as well as the correctly specified GLM. Finally, Web Figure 1 of Web Appendix D plots the variable inclusion proportions for the two BART configurations to examine variable importance. This strategy has been found to work better under a small number of trees, which is reflected in the figure. The only variable used for prediction under M1 was x_{i1}, and BART successfully estimates it as having the largest inclusion proportion. All three variables were used for prediction under M2, and BART estimates relatively similar inclusion proportion values for each variable.

[Figure 2 about here.]

Web Figures 2-3 and Web Tables 2-3 in Web Appendix D summarize the simulation results when the assay accuracies were considered to be unknown and only the DT protocol was implemented. The results appear to be analogous to that of the first simulation configuration that assume assay accuracies are known. Additionally, BART can successfully estimate these

unknown assay accuracies (see Web Table 3). Overall, BART continues to perform well and is robust to the number of trees, even when the assay accuracies are unknown.

5. Iowa Chlamydia Data Analysis

The State Hygienic Laboratory (SHL) at the University of Iowa is the largest public health laboratory in Iowa. Each year, the lab tests thousands of Iowa residents for chlamydia and gonorrhea as part of federally sponsored STD assessment and prevention programs. Individual endocervical swab and urine specimens are collected from various clinics located throughout the state which are then transported to the SHL where group testing is employed. The current SHL screening procedure requires all urine specimens to be individually tested, while a Dorfman testing (DT) protocol is used for swab specimens. That is, swab specimens are tested in master pools, usually of size 4, and the positive master pools are resolved by testing the individual specimen separately. The SHL uses the Aptima Combo 2 Assay (AC2A) to test all collected specimens, both pooled and individual. Pilot data describing the accuracy of the AC2A for individual testing are summarized in the product literature, available at `www.fda.com`; see also Gaydos et al. (2003). We also summarize these pilot data in Web Table 4 of Web Appendix E.

To illustrate our proposed BART methodology, we will analyze chlamydia testing data that the SHL collected from $N = 13,862$ female individuals during one calendar year. The available data consists of testing responses from 4316 individual urine specimens, 416 individual swab specimens, 2273 swab master pools of size 4, 12 swab master pools of size 3, one swab master pool of size 2, and any retesting responses required to resolve positive swab master pools. Additionally, six covariates considered to be potential risk factors were collected on each individual: age (in years, denoted by x_{i1}), a race indicator ($x_{i2} = 1$ if Caucasian and $x_{i2} = 0$ otherwise), an indicator denoting whether the patient reported a new sexual partner in the last 90 days ($x_{i3} = 1$ if affirmative and $x_{i3} = 0$ otherwise), an

indicator denoting whether the patient reported having multiple sexual partners in the last 90 days ($x_{i4} = 1$ if affirmative and $x_{i4} = 0$ otherwise), an indicator denoting whether the patient reported sexual contact with an STD-positive partner in the previous year ($x_{i5} = 1$ if affirmative and $x_{i5} = 0$ otherwise), and an indicator denoting whether the patient presented symptoms of infection ($x_{i6} = 1$ if affirmative and $x_{i6} = 0$ otherwise). To relate an individual's true chlamydia disease status to their available covariate information, we will consider the following BART model

$$\Phi^{-1}\left(P(\widetilde{Y}_i = 1 \big| \boldsymbol{x}_i)\right) = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k, M_k)$$

under two configurations; namely with K=20 trees (a small number of trees for variable selection) and K=200 trees (a large number of trees for flexible prediction), where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i6})'$, $i = 1, 2, \ldots, 13862$. We elicit prior distributions as described in Section 2.1. Although the AC2A was used for testing on all specimen types, it is important to acknowledge differences in how it may perform when testing swab versus urine specimens (Gaydos et al., 2003), and when testing pooled versus individual specimens. With this in mind, we divide the test responses into L=3 strata, which gives rise to the estimation of 3 sets of assay accuracies: $S_{e(1)}$ and $S_{p(1)}$ for swab specimens tested individually, $S_{e(2)}$ and $S_{p(2)}$ for urine specimens tested individually, and $S_{e(3)}$ and $S_{p(3)}$ for swab specimens tested in pools. For these six parameters, we chose informative Beta priors based on the individual AC2A pilot data; for further details, see to Web Appendix E. For purposes of comparison, we also fit a Bayesian generalized linear model (GLM) using all six covariates, following McMahan et al. (2017).

First, we seek to compare the predictive performance of BART and GLM. To do so, we randomly split the data into training and test sets, where 85% of the data was used to train the model and the remaining 15% was allocated to the test set. Note that the true responses (individual chlamydia statuses) are obscured by the assay testing errors. Therefore, it is

not appropriate to conduct an ROC curve analysis as we did for the simulated settings in Section 4. Instead, we will examine the predictive error through the log-likelihood. This is analogous to comparing model-based performance using the cross-entropy loss function Hastie et al. (2009). For each model fit, we use the posterior mean parameter estimates to compute the log-likelihood as a measure of overall fit. Table 2 reports the calculated log-likelihood measures for both in- and out-of-sample. The BART models resulted in larger log-likelihood measures than that of GLM, suggesting that BART fits the data better than GLM and that GLM's parameter estimates are potentially suffering from a misspecification of the mean structure.

To explore this further, the posterior mean estimated functions and the corresponding estimated probabilities from the two BART configurations were plotted against the age covariate x_{i1} for all 32 risk profiles. Web Figure 4 in Web Appendix F displays the results. For a more focused comparison, Figure 3 displays the estimation results for three specific risk profiles. This figure showcases the nonlinear effect of age, and there appears to be an interaction effect among the multiple predictors for these risk profiles. In particular, for patients falling into the risk profile displayed by the blue curve, their risk of infection is larger than the other two profiles until approximately 25 years of age where the risk of infection for patients falling into the profile displayed by the orange curve starts to become larger than the other two risk profiles as age continues to increase. Web Figures 5-7 in Web Appendix F provide additional evidence of nonlinear interactions throughout other risk profiles. Without having to explicitly specify the functional form, BART was able to capture the nonlinear effect of age and any nonlinear, high-order interactions among the multiple risk factors. Note that the two BART configurations yield the same conclusions, but the configuration with a larger number of trees produces approximately smoother estimates.

[Table 2 about here.]

[Figure 3 about here.]

Web Figure 8 in Web Appendix F plots the variable inclusion proportions, which allows us to rank the risk factors by their relative importance in the prediction of disease. While the inclusion proportions are plotted for both BART configurations, examination of variable importance is more effective for a smaller number of trees, as discussed in Section 3.3. For the configuration with K=20 trees, Web Figure 8 insinuates that age is potentially the most influential risk factor, while having multiple sexual partners is ranked the lowest in terms of its relative importance. Finally, Web Table 5 in Web Appendix F provides the posterior mean estimates, the estimated posterior standard deviations, and the 95% equal-tail credible intervals for the six assay accuracies corresponding to the L=3 strata. The amount of variability in the sensitivity estimates for GLM is notably larger than that of BART, likely due to GLM's misspecification of the mean structure.

6. Discussion

BART is an attractive approach for developing flexible predictive models and, in particular, it offers the ability to provide uncertainty in model estimates. In this article, we have developed a general Bayesian additive regression trees (BART) regression technique for potentially misclassified group testing data and individual-level covariate information. The proposed method expands on the methodology described in McMahan et al. (2017) to allow for a more flexible estimation framework that has the ability to handle nonlinear main effects and high-order interaction effects without any input from the researcher. BART also has the ability to assess variable importance by examining the relative frequencies with which the covariates are used as splitting variables in the posterior samples of the K regression trees.

Our proposed BART approach inspires the exploration of other advanced machine learning techniques that could be used for estimation in the group testing setting. There are several

modeling extensions that could be of interest. One possible extension would be the development of regression techniques used to analyze data that incorporates the testing outcomes from multiplex assays; i.e., assays that test specimens for multiple diseases simultaneously. Another useful modeling extension involves the inclusion of the 'dilution effect', a common concern arising in group testing procedures. This dilution effect occurs if the signal from a positive individual's specimen is diluted past an assay's threshold of detection when it is pooled with multiple negative specimens.

References

Abdulhamid, B., Bilder, C., McCutchen, E., Hinrichs, S., Koepsell, S., and Iwen, P. (2020). Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *American Journal of Clinical Pathology* **153,** 715–718.

Abdurrahman, S., Mbanaso, O., Lawson, L., Oladimeji, O., Blakiston, M., Obasanya, J., Dacombe, R., Adams, E., Emenyonu, N., Sahu, S., Creswell, J., and Cuevas, L. (2015). Testing pooled sputum with Xpert MTB/RIF for diagnosis of pulmonary tuberculosis to increase affordability in low-income countries. *Journal of Clinical Microbiology* **53,** 2502–2508.

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

Bleich, J., Kapelner, A., George, E., and Jensen, S. (2014). Variable selection for BART: An application to gene regulation. *The Annals of Applied Statistics* **8,** 1750–1781.

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

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

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.

Delaigle, A. and Tan, R. (2023). Group testing regression analysis with covariates and specimens subject to missingness. *Statistics in Medicine* **42,** 731–744.

Gastwirth, J. and Johnson, W. (1994). Screening with cost effective quality control: Potential applications to hiv and drug testing. *Journal of the American Statistical Association* **89,** 972–981.

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.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2 edition.

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

Johnson, W. and Gastwirth, J. (2000). Dual group screening. *Journal of Statistical Planning and Inferences* **83,** 449–473.

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

Kapelner, A. and Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive

regression trees. *Journal of Statistical Software* **70,** 1–40.

Krajden, M., Cook, D., Mak, A., Chu, K., Chahil, N., Steinberg, M., Rekart, M., and Gilbert, M. (2014). Pooled nucleic acid testing increases the diagnositc yield of acute hiv infections in 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 increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae. *Sexually Transmitted Diseases* **39,** 46–48.

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

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

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.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software* **97,** 1–66.

Torres, I., Albert, E., and Navarro, D. (2020). Pooling of nasopharyngeal swab specimens for SARS-CoV-2 detection by RT-PCR. *Journal of Medical Virology* **92,** 2306–2307.

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 and serum samples. *Biometrics* **56,** 1126–1133.

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.

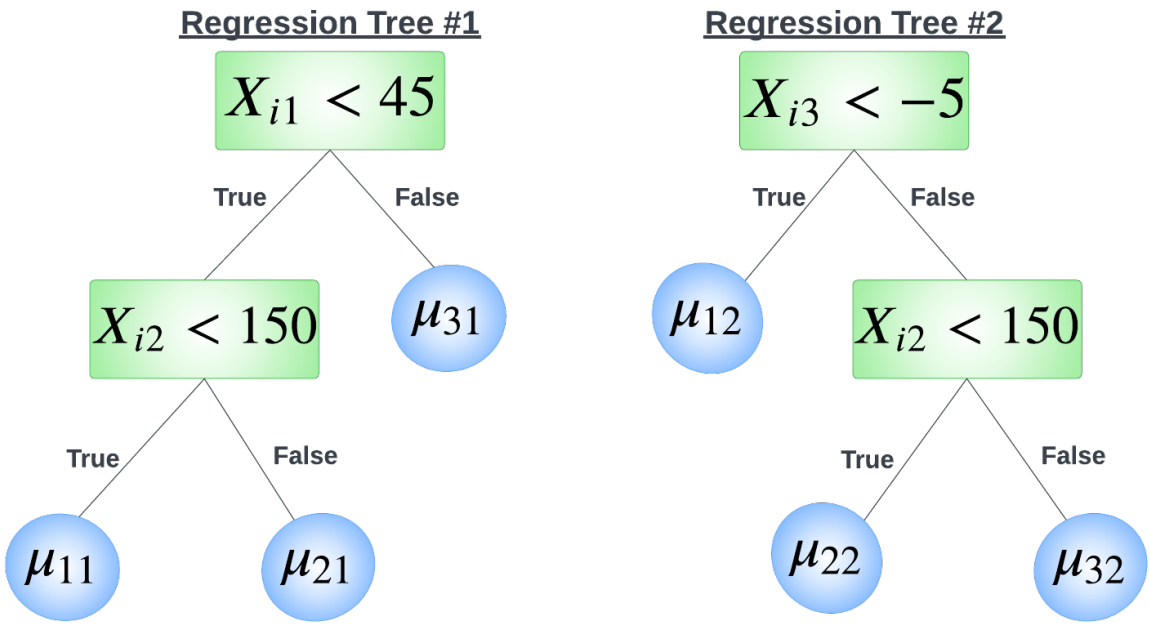

Figure 1. Illustrating the sum of regression trees using a simple two regression tree example.

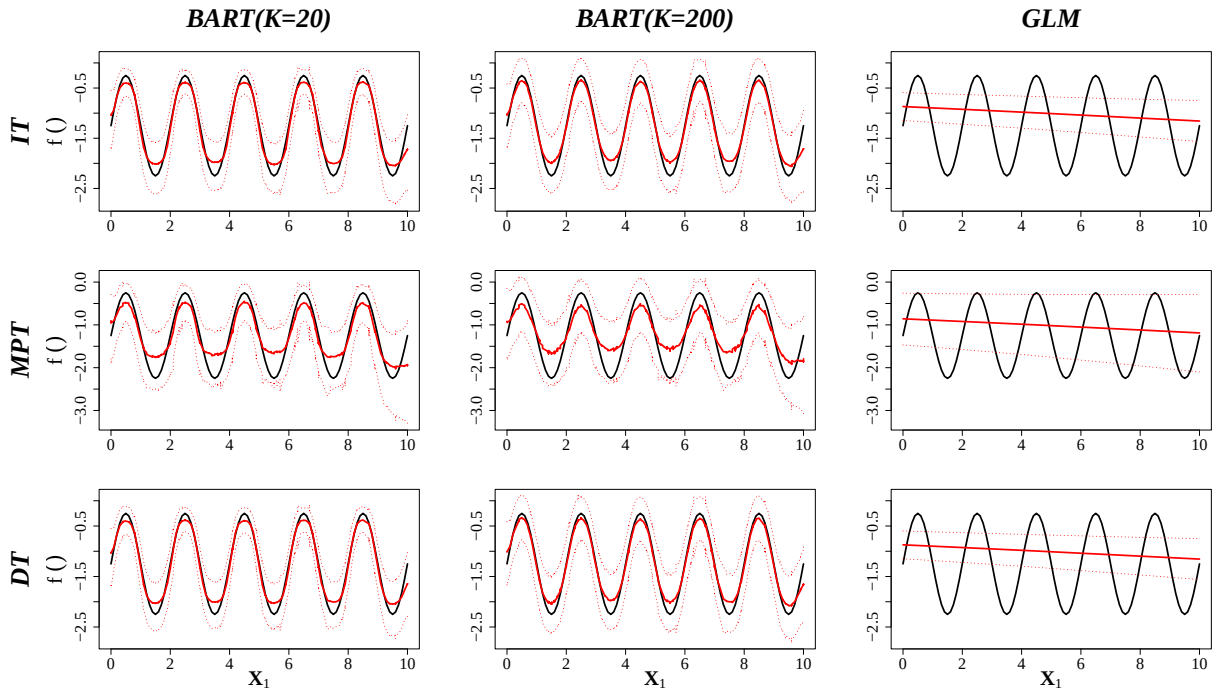

Figure 2. In-sample simulation results for the three model configurations when assay accuracy probabilities are **known**: BART with K=20 trees (left), BART with K=200 trees (middle), and GLM (right) under the group testing protocols IT (top row), MPT (middle row), and DT (bottom row). The black solid curve in each subfigure is the true function $f(\cdot)$ in population model M1. The following are displayed as red curves: the average of the 500 posterior mean estimates (solid curves) and the .025 and .975 posterior mean quantiles (dashed curves).

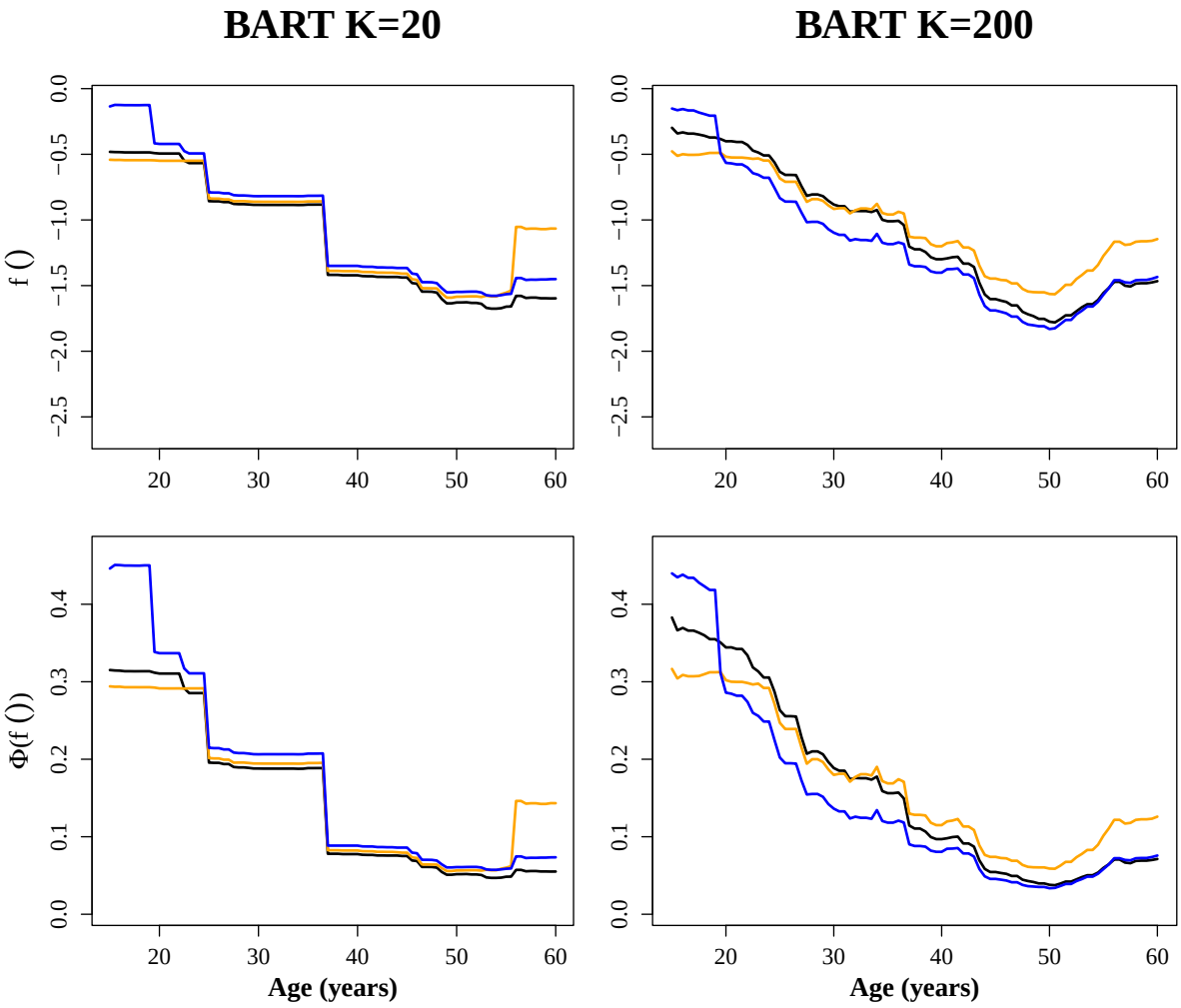

Figure 3. Iowa Chlamydia Data. Posterior mean estimates of the function $f(\cdot)$ (top row) and posterior mean estimates of the probabilities $\Phi(f(\cdot))$ (bottom row) from the BART configurations with K=20 trees (left) and K=200 trees (right), plotted against the age covariate x_{i1} for three risk profiles: non-Caucasian patients that reported sexual contact with an STD-positive partner (black curve); Caucasian patients that reported having a new sexual partner and sexual contact with an STD-positive partner (orange curve); non-Caucasian patients that reported having a new sexual partner, sexual contact with an STD-positive partner, and presented symptoms of infection (blue curve).

Table 1
The values of $\sum_{k=1}^{2} g(\boldsymbol{x}_i; T_k, M_k)$ *from the regression trees in Figure 1.*

i	x_{i1}	x_{i2}	x_{i3}	$g(\boldsymbol{x}_i; T_1, M_1)$	$g(\boldsymbol{x}_i; T_2, M_2)$	$\sum_{k=1}^{2} g(\boldsymbol{x}_i; T_k, M_k)$
1	56	110	-13	μ_{31}	μ_{12}	$\mu_{31} + \mu_{12}$
2	27	173	-3	μ_{21}	μ_{32}	$\mu_{21} + \mu_{32}$
3	41	94	5	μ_{11}	μ_{22}	$\mu_{11} + \mu_{22}$
4	30	213	-9	μ_{21}	μ_{12}	$\mu_{21} + \mu_{12}$
5	48	168	39	μ_{31}	μ_{32}	$\mu_{31} + \mu_{32}$

Table 2

Iowa Chlamydia Data. In- and out-of-sample log likelihood calculated with posterior mean estimates of the assay accuracy probabilities (sensitivity and specificity) and the individual probabilities of being truly positive for chlamydia.

	BART(K=20)	BART(K=200)	GLM
In-Sample	-3329.85	-3320.62	-4379.05
Out-of-Sample	-595.75	-594.95	-802.07