Journal of Machine Learning Research 23 (2024) 1-20 Submitted 1/21; Revised 5/22; Published 9/22

Gradient boosting for group testing

Erica M. Porter EMPORTE@CLEMSON.EDU
School of Mathematical & Statistical Sciences
Clemson University
Clemson, SC 29634, USA

Christopher S. McMahan MCMAHA2@CLEMSON.EDU
School of Mathematical & Statistical Sciences
Clemson University
Clemson, SC 29634, USA

Joshua M. Tebbs TEBBS@STAT.SC.EDU
Department of Statistics
University of South Carolina
Columbia, SC 29208, USA

Christopher R. Bilder BILDER@UNL.EDU
Department of Statistics
University of Nebraska-Lincoln
Lincoln, NE 68583, USA

Editor:

Abstract

When conducting disease screening within a population, it is often more efficient and cost effective to group individual specimens and test them in pools, rather than testing each specimen individually. This method, which is known as group testing, speeds up the diagnostic process and reduces costs, especially when the outcome of interest is rare. However, the data collected from group testing is inherently complex, especially in the presence of imperfect testing, and this complexity can hinder surveillance efforts. To overcome this challenge, we propose a gradient boosting method specifically designed to build predictive models based on group testing data using individual-level predictors. Our framework is flexible, supporting a wide range of weak learners, including regression trees, kernel smoothing, and splines. Our model accommodates data arising from any group testing protocol and accounts for the effects of imperfect testing. To optimize model performance, we develop a cross-validation approach that selects optimal tuning parameters for the weak learners. We explore the performance of our approach through numerical studies. Finally, we apply our method to chlamydia group testing data collected by the State Hygienic Laboratory in Iowa.

Keywords: Pooled testing, gradient boosting, cross-validation, data likelihood, binary classification

1 Introduction

Governments and health care systems are tasked with managing a plethora of health care crises, including infectious disease outbreaks and pandemics, environmental pollution, food-

borne illness, antimicrobial resistance, and health inequalities in underserved communities. In responding to these various crises, diagnostic testing plays a crucial role in identifying, managing, and mitigating their impact. For example, during disease outbreaks, rapid and accurate diagnostic testing helps identify cases early, trace contacts, and implement isolation measures to contain the spread. Similarly, targeted diagnostic testing in underserved communities helps identify disparities in health outcomes, inform resource allocation, and improve access to timely healthcare services. In either case, key limitations to the implementation of large scale screening programs is the cost of diagnostic testing or lack of testing resources; e.g., reagents, test strips, vials, etc.

Dorfman (1943) introduced a seminal solution to these challenges by pioneering a diagnostic strategy known as group (or pool) testing. This approach involves testing pooled samples created by combining biospecimens from multiple individuals. If a pooled sample tests negatively, all individuals contributing to that pool are considered negative. However, if the pool tests positively, additional testing is necessary to identify (decode) which specific individual(s) within the pool is (are) positive; for discussion on decoding strategies, see Kim et al. (2007). In low prevalence settings, group testing can drastically reduce the testing cost associated with classifying all individuals as diseased or not. Given these benefits, group testing has been used to address diagnostic testing needs surrounding numerous public health issues, including vector-borne disease surveillance (Speybroeck et al., 2012), screening during the COVID-19 pandemic (Yu et al., 2021; Eberhardt et al., 2020; Torres et al., 2020; Abdalhamid et al., 2020), detecting food-borne pathogens (Jassem et al., 2016), and screening for sexually transmitted diseases (Gastwirth and Johnson, 1994; Krajden et al., 2014).

In many of these applications, researchers and public health officials are challenged with the complementary tasks of case identification and surveillance, with the latter being the focus of our work here. Specifically, we consider the problem of developing an innovative modeling technique tailored for analyzing data arising from group testing. Due to the effects of imperfect testing and different pooling designs, the analysis of group testing data is a non-trivial task, because the true infection statuses of the individuals are obscured. Existing regression methods for group testing data using generalized linear models (Farrington, 1992; McMahan et al., 2017), mixed effects models (Joyner et al., 2020; Chen et al., 2009), and penalized methods (Gregory et al., 2018) are available. However, these methods require parametric assumptions and have not been used for detection of nonlinear relationships and/or complex interaction effects. Others have developed semiparametric (Wang et al., 2014; Delaigle et al., 2014), nonparametric (Delaigle and Meister, 2011), and additive models (Liu et al., 2020) to allow for nonlinear effects with group testing data. However, these approaches predominantly focus on nonlinear relationships in a single predictor and cannot automatically detect complex interactions.

To overcome these limitations, we propose a gradient boosting methodology that can analyze data arising from any group testing protocol. Gradient boosting is a powerful machine learning technique that offers several advantages over traditional methods, including predictive accuracy, flexibility, incremental learning, feature importance, and others (Mayr et al., 2014; Zhang et al., 2017). Our methods work to create a final model in a stagewise fashion via an ensemble of models, where each additional model (or weak learner) added to the ensemble improves on the limitations of the previous ensemble. By far the most popu-

lar weak learners are tree models, which when coupled with our general gradient boosting approach, allow an end user to aptly detect and account for complex nonlinear relationships and interaction effects.

Of course, any statistical model can be used as a weak learner, and the advantages and disadvantages are application-specific. With the aforementioned advantages, gradient boosting algorithms have been developed for a variety of data structures, including univariate and multivariate regression (Hadiji et al., 2015), binary and ordinal classification (Riccardi et al., 2014), survival endpoints and censored data (He et al., 2015; Zhang et al., 2020; Li et al., 2022), time series (Körner et al., 2018; Nakagawa and Yoshida, 2022), and imaging (Iranzad et al., 2022; Lawrence et al., 2004). In this paper, we seek to extend gradient boosting to allow for the analysis of group testing data. Our approach allows an end user to select any weak learner, thereby providing a versatile analysis framework. The motivation for this work stems from infectious disease screening practices implemented by the State Hygienic Laboratory (SHL) at the University of Iowa. The SHL uses a variant of Dorfman's original testing protocol to test individual subjects for chlamydia. The adoption of this protocol has led to considerable cost savings for the laboratory, but also results in highly complex group testing data that can hinder surveillance efforts.

The remainder of this article is organized as follows. Section 2.1 describes our modeling assumptions, presents the observed data likelihood, and provides details on the proposed gradient boosting algorithm for group testing. Section 3 provides general implementation details, including a cross-validation procedure appropriate for group testing data. Section 4 reports the results of simulation studies to assess the performance of our approach. Section 5 presents an analysis of chlamydia group testing data collected by the SHL. Section 6 concludes with a summary discussion.

2 Methodology

2.1 Notation and Preliminaries

Consider a setting in which N individuals are screened for an infectious agent (e.g., chlamydia, HIV, etc.) by a group testing protocol. Let $\widetilde{Y}_i$ denote the true infection status of the ith individual, for $i = 1, ..., N$, with $\widetilde{Y}_i = 1$ indicating the individual is truly positive and $\widetilde{Y}_i = 0$ otherwise. Furthermore, let $\mathbf{x}_i$ denote a p-dimensional vector of predictor values (e.g., age, race, sex, sexual history, etc.) taken on the ith individual. To relate these variables, we assume that $\widetilde{Y}_i|\mathbf{x}_i \sim \text{Bernoulli}\{H(\mathbf{x}_i)\}$, where $H(\cdot) = [1+\exp\{-F(\cdot)\}]^{-1}$ and $F(\cdot)$ is a unknown function. This formulation captures traditional binary regression models by taking $F(\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of regression coefficients, as well as more modern techniques which are elucidated below. In either case, it is important to note that due to the effects of imperfect testing, the individuals' true infection statuses (i.e., the $\widetilde{Y}_i$'s) are not observed and failing to account for this feature can compromise inference.

The observed data arising from implementing a group testing protocol can be complex. First, there are many protocols available (Dorfman, 1943; Phatarfod and Sudbury, 1994; Kim et al., 2007; Kim and Hudgens, 2009) and most require individuals to be tested in multiple (possibly overlapping) pools and may even further mandate confirmatory testing for quality control purposes (Gastwirth and Johnson, 1994; Johnson and Gastwirth, 2000; Krajden et al., 2014). Thus, to provide a general framework which can incorporate data

from any group testing protocol, we define the index set $\mathcal{P}_j \subset \{1, ..., N\}$ which identifies the individuals contributing to the jth pool, for $j = 1, ..., J$. Let $\widetilde{Z}_j$ denote the true status of the jth pool. We adopt the convention that a pool is positive ($\widetilde{Z}_j = 1$) if it contains at least one infected individual and negative otherwise ($\widetilde{Z}_j = 0$); i.e., $\widetilde{Z}_j = I(\sum_{i \in \mathcal{P}_j} \widetilde{Y}_i > 0)$, where $I(\cdot)$ is the indicator function. Like the individuals' true statuses, the $\widetilde{Z}_j$'s are also unobserved due to the effect of imperfect testing. Instead, we observe the testing response Z_j which can be viewed as an error-contaminated version of $\widetilde{Z}_j$, with $Z_j = 1$ indicating the jth pool tests positively and $Z_j = 0$ otherwise. To quantify the effect of imperfect testing, let $S_{ej} = P(Z_j = 1 \mid \widetilde{Z}_j = 1)$ and $S_{pj} = P(Z_j = 0 \mid \widetilde{Z}_j = 0)$ denote the sensitivity and specificity, respectively, of the assay used to test the jth pool. We allow S_{ej} and S_{pj} to be pool specific, thus allowing for the use of multiple assays and/or the effect that pool size (i.e., the cardinality of $\mathcal{P}_j$) may have on an assay's performance.

To relate the individual-level model to the observed testing responses $\mathbf{Z} = (Z_1, ..., Z_J)'$, we assume the responses in $\mathbf{Z}$ are conditionally independent given the true statuses $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, ..., \widetilde{Z}_J)'$ and the conditional distribution $\mathbf{Z} \mid \widetilde{\mathbf{Z}}$ does not depend on the predictor variables. Under these assumptions, the observed data likelihood is

$$L(\mathbf{Z} \mid F) = \sum_{\widetilde{\mathbf{Y}} \in \boldsymbol{\mathcal{Y}}} \left[\prod_{j=1}^{J} \delta(\widetilde{Z}_j, Z_j) \prod_{i=1}^{N} H(\mathbf{x}_i)^{\widetilde{Y}_i} \{1 - H(\mathbf{x}_i)\}^{1-\widetilde{Y}_i} \right],$$

where $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_1, ..., \widetilde{Y}_N)'$ is the vector of true disease statuses, $\boldsymbol{\mathcal{Y}}$ is the support of $\widetilde{\boldsymbol{Y}}$, and $\delta(\widetilde{Z}_j, Z_j) = S_{ej}^{Z_j \widetilde{Z}_j}(1 - S_{ej})^{(1-Z_j)\widetilde{Z}_j}(1 - S_{pj})^{Z_j(1-\widetilde{Z}_j)} S_{pj}^{(1-Z_j)(1-\widetilde{Z}_j)}$. Note that the form of $L(\mathbf{Z} \mid F)$ is completely general and therefore applicable for any protocol which uses group testing. In Section 2.2, we examine how gradient boosting can be implemented when the master pools are tested and no further retesting occurs (Vansteelandt et al., 2000; Delaigle and Meister, 2011; Delaigle et al., 2014). Although case identification is not the goal, this protocol serves to provide a straightforward introduction to gradient boosting with group testing. In Section 2.3, we then develop a more general gradient boosting algorithm that allows for data arising from more complex group testing protocols where both case identification and surveillance are simultaneous goals.

2.2 Gradient boosting for master pool testing

Under master pool testing, each individual is assigned to exactly one pool which is subsequently tested, and, regardless of the test outcome, no follow-up testing is performed. Under this protocol, $L(\mathbf{Z} \mid F)$ simplifies substantially and the log-likelihood of the observed data is

$$l_M(\mathbf{Z} \mid F) = \sum_{j=1}^{J} \{Z_j \log(q_j) + (1 - Z_j) \log(1 - q_j)\}, \tag{1}$$

where $q_j = S_{ej} + (1 - S_{ej} - S_{pj}) \prod_{i \in \mathcal{P}_j} \{1 - H(\mathbf{x}_i)\}$. From here, we treat the negative of (1) as the loss function and proceed to develop our gradient boosting algorithm. In general, at the $(m+1)$th step of the algorithm we first compute the pseudo-residuals, given the current

model fit, which in the case of master pool testing are given by

$$R_i^{(m)} = \left.\frac{\partial l_M(\mathbf{Z} \mid F)}{\partial F(\mathbf{x}_i)}\right|_{F(\cdot)=F^{(m)}(\cdot)} = \frac{(Z_j - q_j^{(m)})q_{j(i)}^{(m)}}{q_j^{(m)}(1-q_j^{(m)})}, \tag{2}$$

where $F^{(m)}(\cdot)$ denotes the current estimate of $F(\cdot)$, and

$$q_j^{(m)} = S_{ej} + (1 - S_{ej} - S_{pj}) \prod_{i \in \mathcal{P}_j} \{1 - H^{(m)}(\mathbf{x}_i)\}$$

$$q_{j(i)}^{(m)} = -\left[(1 - S_{ej} - S_{pj}) \prod_{i \in \mathcal{P}_j} \{1 - H^{(m)}(\mathbf{x}_i)\}\right] H^{(m)}(\mathbf{x}_i).$$

In the expressions above, we adopt the convention that $H^{(m)}(\cdot)$ is obtained by replacing $F(\cdot)$ in $H(\cdot)$ with the current estimate $F^{(m)}(\cdot)$. Once the residuals are computed, a weak learner (e.g., linear model, splines, regression trees, etc.) is fit to $\{(R_i^{(m)}, \mathbf{x}_i), i = 1, ..., n\}$, treating $R_i^{(m)}$ as the response variable and $\mathbf{x}_i$ as the predictor variables. Denote this model fit by $w^{(m)}(\cdot)$. Based on this weak learner, we then update the model fit as

$$F^{(m+1)}(\cdot) = F^{(m)}(\cdot) + \gamma^{(m)} w^{(m)}(\cdot),$$

where $\gamma^{(m)}$ is a learning rate. Depending on the weak learner being implemented, the learning rate can be determined in various ways; e.g., user-specified, via a line search, component-specific (tree-based learners), etc.

2.3 General gradient boosting

We now consider the more challenging setting where individuals are tested in multiple pools as part of a group testing protocol for case identification which may or may not include additional retests for quality control. To present our approach, we first decompose the observed data likelihood by dividing the individuals under study into K non-overlapping subgroups. Ideally, these subgroups should be constructed with two considerations in mind. First, they should be limited in size with respect to the number of individuals as larger subgroup sizes increase computational complexity. Second, subgroups should be formed with individuals from a common subgroup. For many group testing protocols, these subgroups arise naturally; e.g., for Dorfman testing, the master pool serves as a subgroup. For other protocols, the notion of a subgroup may not be as obvious. In array testing (Farrington, 1992; Kim et al., 2007; Kim and Hudgens, 2009), for example, the subgroup is the entire array, rather than initial row and column pools within the array.

Define the index sets B_k and C_k to track the tests and individuals associated with the kth subgroup. Our specifications above require

1. $\cup_{k=1}^{K} B_k = \{1, ..., J\}$ and $B_k \cap B_{k'} = \emptyset$ for all $k \neq k'$
2. $\cup_{j \in B_k} \mathcal{P}_j = C_k$ and $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$.

With these sets defined, we can rewrite the observed data likelihood as

$$L(\mathbf{Z} \mid F) = \prod_{k=1}^{K} \sum_{\widetilde{\mathbf{Y}}_k \in \boldsymbol{\mathcal{Y}}_k} \left[\prod_{j \in B_k} \delta(\widetilde{Z}_j, Z_j) \prod_{i \in C_k} H(\mathbf{x}_i)^{\widetilde{Y}_i} \{1 - H(\mathbf{x}_i)\}^{1-\widetilde{Y}_i} \right], \tag{3}$$

where $\widetilde{\mathbf{Y}}_k = \{\widetilde{Y}_i : i \in C_k\}$ and $\boldsymbol{\mathcal{Y}}_k$ is the support of the binary vector $\widetilde{\mathbf{Y}}_k$. Again treating the negative of the log-likelihood as a loss function, we develop our gradient boosting algorithm by identifying the pseudo-residuals

$$R_i^{(m)} = \frac{\sum_{\widetilde{\mathbf{Y}}_k \in \boldsymbol{\mathcal{Y}}_k} \{\widetilde{Y}_i - H^{(m)}(\mathbf{x}_i)\} \prod_{j \in B_k} \delta(\widetilde{Z}_j, Z_j) \prod_{i \in C_k} H^{(m)}(\mathbf{x}_i)^{\widetilde{Y}_i} \{1 - H^{(m)}(\mathbf{x}_i)\}^{1-\widetilde{Y}_i}}{\sum_{\widetilde{\mathbf{Y}}_k \in \boldsymbol{\mathcal{Y}}_k} \prod_{j \in B_k} \delta(\widetilde{Z}_j, Z_j) \prod_{i \in C_k} H^{(m)}(\mathbf{x}_i)^{\widetilde{Y}_i} \{1 - H^{(m)}(\mathbf{x}_i)\}^{1-\widetilde{Y}_i}},$$

for $i \in C_k$. Model fitting then proceeds as described in Section 2.2. That is, a weak learner is chosen and fit to $\{(R_i^{(m)}, \mathbf{x}_i), i = 1, ..., n\}$, treating $R_i^{(m)}$ as the response variable and $\mathbf{x}_i$ as the predictors. Model updates are determined analogously.

3 Implementation

3.1 Cross-validation for tuning parameter selection

In implementing a gradient boosting algorithm, it is necessary to select tuning parameters for each of the weak learners. For example, when fitting regression trees common tuning parameters include the maximum tree depth and the minimum number of observations in a terminal node. For kernel smoothing we tune the bandwidth, and for splines we select the number of knots used to define the spline basis. Further, a user must also select the number of epochs (boosting iterations) used to fit the model. Traditionally, these choices are guided by S-fold cross validation, where the tuning parameters/epochs are chosen to minimize various performance metrics; e.g., mean squared error or cross entropy loss. However, these traditional methods are not applicable in the considered setting since the true individual disease statuses are not observed.

In what follows, we propose a S-fold cross validation strategy which makes use of an evaluation metric that is inspired by the cross entropy loss. This approach has several nuances that arise due to the complex structure of group testing data. First, unlike traditional cross validation strategies which randomly assign each individual/observation to one of S folds, our approach assigns entire subgroups of individuals to the various folds. Proceeding in this fashion is necessary, since computing the pseudo-residual $R_i^{(m)}$, for $i \in C_k$, requires information on all other individuals in C_k. Second, given that the individuals true statuses are not observed, our approach uses the log-likelihood of the testing outcomes as a performance metric to guide tuning parameter selection.

In particular, our cross validation procedure proceeds as follows. First, we define a grid of tuning parameters and let $\omega \in \{1, \ldots \Omega\}$ index this grid. For example, if using regression trees as the weak learner, ω might represent particular values for the maximum tree depth, minimum number of observations in a terminal node, and the number of epochs used to complete model fitting. Second, we randomly assign each subgroup to one of the S folds and define the index set $A_s \subset \{1, ..., K\}$ which identifies the subgroups assigned to the sth

fold, for $s = 1, ..., S$. Following the usual tenets of cross validation, we hold one fold out at a time for testing and use the remaining folds as training data. Let $\widehat{F}_{\omega,s}$ denote the model fit using tuning parameters ω holding the sth fold out. Based on these estimates, our evaluation metric is given by $CV_\omega = \sum_{s=1}^{S} \log\{L_{\omega,s}(\mathbf{Z}_s \mid \widehat{F}_{\omega,s})\}$, where

$$L_{\omega,s}(\mathbf{Z}_s \mid \widehat{F}_{\omega,s}) = \prod_{k \in A_s} \sum_{\widetilde{\mathbf{Y}}_k \in \boldsymbol{\mathcal{Y}}_k} \left[\prod_{j \in B_k} \delta(\widetilde{Z}_j, Z_j) \prod_{i \in C_k} \widehat{H}_{\omega,s}(\mathbf{x}_i)^{\widetilde{Y}_i} \{1 - \widehat{H}_{\omega,s}(\mathbf{x}_i)\}^{1-\widetilde{Y}_i} \right],$$

$\widehat{H}_{\omega,s}(\cdot)$ is obtained by replacing $F(\cdot)$ in $H(\cdot)$ with $\widehat{F}_{\omega,s}$, and $\mathbf{Z}_s = \{Z_j : j \in \cup_{k \in A_s} B_k\}$. Based on this metric, we select the tuning configuration ω that results in the largest CV_ω. Once complete, we train the gradient boosting model on the entire data set under this tuning configuration to obtain our estimate of F.

3.2 Learning Rate

Another tuning parameter that is often selected for gradient boosting via cross-validation is the learning rate. There are many ways to specify the learning rate in gradient boosting algorithms. For example, a user can fix an overall learning rate γ, specify $\gamma^{(m)}$ at each epoch, or select an overall γ using a grid search. Through numerical studies, we have found that it is computationally intensive to conduct a grid search for the purposes of identifying a global γ to be used to fit the final model. Further, for the settings studied here, we have found that the gradient boosting algorithm performs better when the learning rate is optimally determined at each epoch. That is, herein we identify the learning rate at each epoch as

$$\gamma^{(m)} = \arg\max_{\gamma} L\{\mathbf{Z} | F^{(m)}(\cdot) + \gamma w^{(m)}(\cdot)\}.$$

Experience has shown that proceeding in this fashion quickly improves the model fit and decreases computation by substantially reducing the number of required epochs.

3.3 Variable Importance

Variable importance, the relative ability of available predictors to predict the response, is often of interest in machine learning applications. Many variable importance measures have been proposed that are specific to the model type or algorithm being used. We develop a versatile framework for gradient boosting for group-testing data that accommodates many types of weak learners. We also select a different learning rate at each epoch via optimization, so that the model estimate and contribution of each predictor variable are not weighted equally across the epochs. Therefore, we adopt an algorithm-agnostic variable importance measure developed by Williamson et al. (2023) to provide relative measures of predictive ability for predictor variables used in our gradient boosting approach.

For a prediction function F, Williamson et al. (2023) describe the deviance predictiveness measure $V(F)$. $V(F)$ is defined as 1 minus the ratio of the log-likelihood computed under a model fit F and the log-likelihood of the null model. Thus, for group-testing data, the deviance predictiveness measure for a model F is:

$$V(F) = 1 - \frac{\log[L(\mathbf{Z}|F)]}{\log[L(\mathbf{Z}|F_0)]},$$

where F_0 is the null model. $V(F)$ measures the increase in information that results from using the predictors in F to predict the response instead of using only the null model. Based on this measure, an estimator of the importance of predictor x_j relative to all others is:

$$\psi_{n,j} = V(\widehat{F}) - V(\widehat{F}_{-j}), \tag{4}$$

where $\widehat{F}$ denotes the final gradient-boosted model containing all available predictors and $\widehat{F}_{-j}$ is the final gradient boosted model obtained from all predictor variables except x_j. The quantity $\psi_{n,j}$ estimates the increase in the deviance predictiveness measure when predictor x_j is removed, compared to the full model. Thus, larger values of $\psi_{n,j}$ indicate greater variable importance for x_j among the predictors considered in the full model.

4 Simulation Results

4.1 Example 1

To evaluate the performance of our gradient boosting method for group testing data, and to demonstrate its versatility with regards to weak learner selection, we first study the following setting relating a single predictor variable to the true disease statuses. The true disease statuses are distributed $\widetilde{Y}_i|x_i \sim \text{Bernoulli}\{H(x_i)\}$, where $H(\cdot) = [1+\exp\{-F(\cdot)\}]^{-1}$, and

$$F(x_i) = -2 + 0.5\sin(x_i), \tag{5}$$

where $x_i \sim \text{Unif}(0, 2\pi),\ i = 1, \ldots, N$. This setting produces an average prevalence rate of 12%, close to that of the study presented in Section 5. To create group testing data, we first use the aforementioned model to randomly generate $N = 10000$ individual true statuses and then we simulate the screening of these individuals under Dorfman's testing protocol. Briefly, this protocol proceeds to assign individuals to (master) pools, which are subsequently tested. If a pool tests negative, no further testing is done. However, if a pool tests positive, each individual in that pool is retested individually. In our implementation, we consider master pools of size 4 and simulate testing outcomes that are subject to diagnostic error. That is, we simulate the $Z_j \sim \text{Bernoulli}\{S_{ej}\widetilde{Z}_j + (1 - S_{pj})(1 - \widetilde{Z}_j)\}$, where $\widetilde{Z}_j = I(\sum_{i\in\mathcal{P}_j} \widetilde{Y}_i > 0)$, $S_{ej} = 0.95$, and $S_{pj} = 0.98$, for $j = 1, ..., J$. We repeat this simulation procedure to obtain 500 group testing data sets.

To demonstrate the performance of our gradient boosting algorithm under different weak learners, we use our approach to analyze the 500 group testing data sets under three such models: regression trees, splines, and kernel smoothing. Each of these modeling choices requires the further specification of tuning parameters. To select the tuning parameter configuration, we made use of the cross-validation strategy outlined in Section 3.1, assigning subgroups to one of 5 folds. For regression trees we tuned over the maximum tree depth, allowing values of 1 and 5, and the minimum number of observations in any terminal node, allowing values of 100, 200, and 300. For splines we tuned over the number of interior knots, applying 1, 4, 7, and 10 knots. For the kernel smoothing approach we obtained

results using bandwidths equal to 0.5, 1, 1.5, and 2. We also tuned across the number of gradient boosting epochs with each weak learner, setting a maximum number of 30 epochs. Note, from preliminary studies, we found that 30 epochs were more than sufficient in this setting when implementing the approach described in Section 3.2 to select the learning rate.

Figure 1 provides a summary of our estimates of F. In particular, for each of the weak learner specifications we provide the pointwise mean, as well as the 0.025 and 0.975 quantiles, of the 500 estimated functions from our simulation. From these results, we see that our proposed gradient boosting algorithm works well. That is, the average estimate of F closely matches the truth, and the 95% quantile curves cover the entire true model. Moreover, there are no appreciable differences in the estimates across the three weak learner specifications, with the exception that the model based on regression trees is slightly less accurate and precise. The findings from this study demonstrates that our gradient boosting approach for group-testing data can successfully be used to model the probability of a low-prevalence disease as a nonlinear function of a predictor variable using a variety of weak learners.

[Figure 1 about here.]

4.2 Example 2

To further evaluate the performance of our gradient boosting method for group testing data, we next consider a setting in which multiple predictors are available. In this study, we simulate the true individual statuses according to the following model

$$F(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{3} \mathrm{g}_j(x_{ij}), \tag{6}$$

where $x_{ij} \sim \text{Unif}(-1, 1)$, for $j = 1, 2, 3$ and $i = 1, ..., N$,

$$\mathrm{g}_1(x_{i1}) = \beta_1 \sin(2\pi x_{i1}), \qquad \mathrm{g}_2(x_{i2}) = \beta_2 x_{i2}, \qquad \mathrm{g}_3(x_{i3}) = \beta_3 x_{i3}^2, \tag{7}$$

and $\boldsymbol{\beta} = (-2, 0.5, 0.5, 0.5)^T$. These functions were chosen to represent a variety of linear and nonlinear effects and the settings for $\boldsymbol{\beta}$ were chosen to provide an average prevalence rate of approximately 15%. We generated data in the same manner as in Example 1, simulating $N = 10000$ true statuses related to the predictors according to (6) and (7), and simulated the testing responses $\mathbf{Z}$ according to Dorfman Testing with master pools of size 4. We generated a total of 500 data sets from this setting.

We applied our gradient boosting method to each of the 500 data sets. For this setting we focus on the use of regression trees as weak learners since it is relatively straightforward to fit trees with multiple predictors. Cross-validation was performed in the same manner as for Example 1 using 5-fold cross-validation on each data set. For each data set, once a tuning parameter configuration was chosen, we obtained estimates of each function in (7) by obtaining gradient boosted estimates with two of the predictors set to 0 at a time.

Figure 2 provides a summary of our estimates of the functions in (7). In particular, we provide the pointwise mean, as well as the 0.025 and 0.975 quantiles, of the 500 estimated functions from our simulation. The findings from this study suggest that our proposed

gradient boosting algorithm works well in the multiple predictor setting. That is, the average estimate of g_j closely matches the truth, and the 95% quantile curves cover the entire true model, for $j = 1, 2, 3$. For a highly nonlinear relationship like the one in Figure 2a, the average estimate is approximately unbiased for most values of the predictor variable, excepting the steepest regions of the curve. However, even for such a highly nonlinear function of the predictor variable, the 95% quantile curves completely capture the underlying model. Relationships such as the linear and curvilinear ones depicted in Figures 2b and 2c are closer to those most commonly expected for group testing data. For these predictor relationships, approximate bias is almost nonexistent in the average function estimate. This suggests that when multiple predictor variables are linked to individual disease statuses in varying ways, our gradient boosting method can (i) accurately estimate the true model and (ii) uncover both linear and nonlinear relationships between the predictors and individual statuses.

[Figure 2 about here.]

5 Iowa Chlamydia Data

We apply our gradient boosting approach to a data set collected at the State Hygienic Laboratory (SHL) at the University of Iowa. The SHL receives thousands of specimens from clinics throughout Iowa to be tested for chlamydia each year. We analyze data collected from $N = 13862$ individual females, consisting of 9546 swab specimens and 4315 urine specimens. The current testing protocol at the SHL is to test swab specimens using Dorfman Testing and to individually test all urine specimens. For this data, tests were performed on the N specimens using 416 individual swab specimens, 2273 swab master pools of size 4, 12 swab master pools of size 3, 1 swab master pool of size 2, and 4316 individual urine specimens. Subsequently, any positive master pool results were resolved by retesting each contributing specimen individually. For swab specimens tested in pools or individually, the test sensitivity and specificity are 0.942 and 0.976, respectively. For urine specimens sensitivity is 0.947 and specificity is 0.989. We consider 5 predictor variables consisting of demographic and patient response data for each individual: age (x_{i1}, in years), a race indicator ($x_{i2} = 1$ if not Caucasian, $x_{i2} = 0$ otherwise), an indicator for whether the individual reported having multiple sexual partners in the last 90 days ($x_{i3} = 1$ if affirmative, $x_{i3} = 0$ otherwise), an indicator for whether the individual reported having sexual contact with an STD-positive partner within the previous year ($x_{i4} = 1$ if affirmative, $x_{i4} = 0$ otherwise), and an indicator for whether the patient exhibited any symptoms of infection ($x_{i5} = 1$ if affirmative, $x_{i5} = 0$ otherwise).

We used our gradient boosting approach to fit the model $\widetilde{Y}_i|\mathbf{x}_i \sim \text{Bernoulli}\{H(\mathbf{x}_i)\}$, where $H(\mathbf{x}_i) = [1 + \exp\{-F(\mathbf{x}_i)\}]^{-1}$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})'$. We used 5-fold cross-validation to select the regression tree tuning parameters used to fit the final model. Cross-validation selected a maximum tree depth of 1 among values of 1, 2, and 3. The selected minimum number of observations in any terminal node was 75, among values 35 and 75. We performed cross-validation for the SHL chlamydia data multiple times, using different configurations for the training and test sets each time. The selected number of epochs was inconsistent, but every cross-validation attempt selected a maximum tree depth of 1 and 75 minimum observations per node. Further exploration at this tree specification showed that

the model stabilizes after 38 epochs when fit on the entire data set. We obtained estimates of $F(\cdot)$ for each of the 16 risk profiles resulting from all possible combinations of the binary predictors. Figure 3 plots the estimates of $F(\cdot)$ across age when only one of the binary predictors is set to 1 at a time.

[Figure 3 about here.]

Figure 3 indicates a nonlinear effect of age on the probability of chlamydia infection. Examining the risk profiles for one indicator variable at a time indicates similar estimates for each of non-Caucasian patients, patients reporting multiple sexual partners in the last 90 days, and patients exhibiting symptoms when age is also accounted for in the model. Cross-validation selected a maximum tree depth of 1, which precludes interaction among the predictor variables in the final model fit. While interactions between the predictor variables were not strong enough for cross-validation to choose a larger tree depth, we were able to obtain estimates of the additive effect of each indicator variable by subtracting the estimated model containing only age from each model in Figure 3. Table 1 lists the baseline log odds of chlamydia infection obtained from the model intercept, and the estimated increase in log odds of infection that each indicator variable provides. In keeping with the results from Figure 3, having contact with an STD-positive partner provides the largest increase in the log odds of infection, while race, having multiple partners, and exhibiting symptoms produce similar additive effect sizes.

[Table 1 about here.]

We also obtained measures of variable importance as described in Section 3.3 for all 5 predictors considered with the SHL chlamydia data. Figure 4 plots values for the variable importance measures scaled between 0 and 100. Age has the highest predictive ability relative to the remaining predictors studied here, followed by contact with an STD-positive partner.

[Figure 4 about here.]

6 Discussion

We have presented a versatile gradient boosting framework for analyzing data arising from group-testing protocols. Our framework provides a nonparametric approach for automatic detection of complex non-linear relationships between the disease status and individual-level predictors using a wide variety of weak learners. We developed both gradient boosting for master pool testing, where no follow-up testing is performed, and for the general case that applies to any group-testing protocol. Our flexible cross-validation approach can tune over a number of weak learner parameters as needed, and reduces computation by optimizing the learning rate across all epochs. We showed through simulation that our approach can accurately detect nonlinear relationships and recover individual additive effects from multiple predictor variables. We applied our gradient boosting approach to a data set where female patients were tested for chlamydia in Iowa. We found a nonlinear effect of patient age on the probability of chlamydia infection and, using established importance metrics,

provided relative measures of predictive ability for several demographic and patient response variables. The code functions and a reproducible example for our proposed method can be found at https://github.com/emporte2/GB4GT.

Acknowledgments and Disclosure of Funding

We thank Jeffrey Benfer and Kristofer Eveland at the State Hygienic Laboratory at University of Iowa. This work was funded by Grant R01 AI121351 from the National Institutes of Health and Grant OIA-1826715 from the National Science Foundation.

References

Baha Abdalhamid, Christopher Bilder, Emily McCutchen, Steven Hinrichs, Scott Koepsell, and Peter Iwen. Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *American Journal of Clinical Pathology*, 153(6):715–718, 2020. doi: 10.1093/ajcp/aqaa064.

Peng Chen, Joshua Tebbs, and Christopher Bilder. Group testing regression models with fixed and random effects. *Biometrics*, 65(4):1270–1278, 2009. doi: 10.1111/j.1541-0420.2008.01183.x.

Aurore Delaigle and Alexander Meister. Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association*, 106(494):640–650, 2011. doi: 10.1198/jasa.2011.tm10520.

Aurore Delaigle, Peter Hall, and Justin Wishart. New approaches to non-and semiparametric regression for univariate and multivariate group testing data. *Biometrika*, 101:567–585, 2014. doi: 10.1093/biomet/asu025.

Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14(4):436–440, 12 1943. doi: 10.1093/ajcp/aqaa064/10.1214/aoms/1177731363.

Jens Eberhardt, Nikolas Breuckmann, and Christiane Eberhardt. Multi-stage group testing improves efficiency of large-scale COVID-19 screening. *Journal of Clinical Virology*, 128:104382, 2020. ISSN 1386-6532. doi: 10.1016/j.jcv.2020.104382.

Paddy Farrington. Estimating prevalence by group testing using generalized linear models. *Statistics in Medicine*, 11(12):1591–1597, 1992. doi: 10.1002/sim.4780111206.

Joseph Gastwirth and Wesley Johnson. Screening with cost-effective quality control: Potential applications to HIV and drug testing. *Journal of the American Statistical Association*, 89(427):972–981, 1994. doi: 10.1080/01621459.1994.10476831.

Karl Gregory, Dewei Wang, and Christopher McMahan. Adaptive elastic net for group testing. *Biometrics*, 75(1):13–23, 09 2018. ISSN 0006-341X. doi: 10.1111/biom.12973.

Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, 100:477–507, 2015. doi: 10.1007/s10994-015-5506-z.

Kevin He, Yanming Li, Ji Zhu, Hongliang Liu, Jeffrey Lee, Christopher Amos, Terry Hyslop, Jiashun Jin, Huazhen Lin, Qinyi Wei, and Yi Li. Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics*, 32(1):50–57, 09 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv517.

Reza Iranzad, Xiao Liu, W Art Chaovalitwongse, Daniel Hippe, Shouyi Wang, Jie Han, Phawis Thammasorn, Chunyan Duan, Jing Zeng, and Stephen Bowen. Gradient boosted trees for spatial data and its application to medical imaging data. *IISE Transactions*

on Healthcare Systems Engineering, 12(3):165–179, 2022. doi: 10.1080/24725579.2021.1995536.

Agatha Jassem, Frank Chou, Cathevine Yang, Matthew Croxen, Katarina Pintar, Ana Paccagnella, Linda Hoang, and Natalie Prystajecky. Pooled nucleic acid amplification test for screening of stool specimens for Shiga toxin-producing Escherichia coli. *Journal of Clinical Microbiology*, 54(11):2711–2715, 2016. doi: 10.1128/JCM.01373-16.

Wesley Johnson and Joseph Gastwirth. Dual group screening. *Journal of Statistical Planning and Inference*, 83(2):449–473, 2000. doi: 10.1016/S0378-3758(99)00100-7.

Chase Joyner, Christopher McMahan, Joshua Tebbs, and Christopher Bilder. From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data. *Biometrics*, 76(3):913–923, 2020. doi: 10.1111/biom.12704.

Hae Kim and Michael Hudgens. Three-dimensional array-based group testing algorithms. *Biometrics*, 65(3):903–910, 2009. doi: 10.1111/j.1541-0420.2008.01158.x.

Hae Kim, Michael Hudgens, Jonathan Dreyfuss, Daniel Westreich, and Christopher Pilcher. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, 63(4):1152–1163, 2007. doi: 10.1111/j.1541-0420.2007.00817.x.

Philipp Körner, Rico Kronenberg, Sandra Genzel, and Christian Bernhofer. Introducing gradient boosting as a universal gap filling tool for meteorological time series. *Meteorologische Zeitschrift*, 27(5):369–376, 2018. doi: 10.1127/metz/2018/0908.

Mel Krajden, Darrel Cook, Annie Mak, Ken Chu, Navdeep Chahil, Malcolm Steinberg, Michael Rekart, and Mark Gilbert. 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(1):132–137, 2014. doi: 10.1016/j.jcv.2014.06.024.

Rick Lawrence, Andrew Bunn, Scott Powell, and Michael Zambon. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90(3):331–336, 2004. doi: 10.1016/j.rse.2004.01.007.

Kaiqiao Li, Sijie Yao, Zhenyu Zhang, Biwei Cao, Christopher Wilson, Denise Kalos, Pei Fen Kuan, Ruoqing Zhu, and Xuefeng Wang. Efficient gradient boosting for prognostic biomarker discovery. *Bioinformatics*, 38(6):1631–1638, 2022. doi: 10.1093/bioinformatics/btab869.

Yan Liu, Christopher McMahan, Joshua Tebbs, Colin Gallagher, and Christopher Bilder. Generalized additive regression for group testing data. *Biostatistics*, 22(4):873–889, 02 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa003.

Andreas Mayr, Harald Binder, Olaf Gefeller, and Matthias Schmid. The evolution of boosting algorithms. *Methods of Information in Medicine*, 53(06):419–427, 2014. doi: 10.3414/ME13-01-0122.

Christopher McMahan, Joshua Tebbs, Timothy Hanson, and Christopher Bilder. Bayesian regression for group testing data. *Biometrics*, 73(4):1443–1452, 2017. doi: 10.1111/biom.12704.

Kei Nakagawa and Kenichi Yoshida. Time-series gradient boosting tree for stock price prediction. *International Journal of Data Mining, Modelling and Management*, 14(2):110–125, 2022. doi: 10.1504/IJDMMM.2022.123357.

Ravindra Phatarfod and Aidan Sudbury. The use of a square array scheme in blood testing. *Statistics in Medicine*, 13(22):2337–2343, 1994. doi: 10.1002/sim.4780132205.

Annalisa Riccardi, Francisco Fernández-Navarro, and Sante Carloni. Cost-sensitive AdaBoost algorithm for ordinal regression based on extreme learning machine. *IEEE Transactions on Cybernetics*, 44(10):1898–1909, 2014. doi: 10.1109/TCYB.2014.2299291.

Niko Speybroeck, Christopher Williams, Kora Lafia, Brecht Devleesschauwer, and Dirk Berkvens. Estimating the prevalence of infections in vector populations using pools of samples. *Medical and Veterinary Entomology*, 26(4):361–371, 2012. doi: 10.1111/j.1365-2915.2012.01015.x.

Ignacio Torres, Eliseo Albert, and David Navarro. Pooling of nasopharyngeal swab specimens for SARS-CoV-2 detection by RT-PCR. *Journal of Medical Virology*, 92(11):2306, 2020. doi: 10.1002/jmv.25971.

Stijn Vansteelandt, Els Goetghebeur, and Thomas Verstraeten. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*, 56(4):1126–1133, 2000. doi: 10.1111/j.0006-341x.2000.01126.x.

Dewei Wang, Christopher McMahan, Colin Gallagher, and Karunarathna Kulasekera. Semiparametric group testing regression models. *Biometrika*, 101(3):587–598, 2014. doi: 10.1093/biomet/asu007.

Brian Williamson, Peter Gilbert, Noah Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023. doi: 10.1080/01621459.2021.2003200.

Jiali Yu, Yiduo Huang, and Zuo-Jun Shen. Optimizing and evaluating PCR-based pooled screening during COVID-19 pandemics. *Scientific Reports*, 11(1):21460, 2021. doi: 10.1038/s41598-021-01065-0.

Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128–150, 2017. ISSN 0957-4174. doi: 10.1016/j.eswa.2017.04.003.

Pingye Zhang, Junshui Ma, Xinqun Chen, and Yue Shentu. A nonparametric method for value function guided subgroup identification via gradient tree boosting for censored survival data. *Statistics in Medicine*, 39(28):4133–4146, 2020. doi: 10.1002/sim.8714.

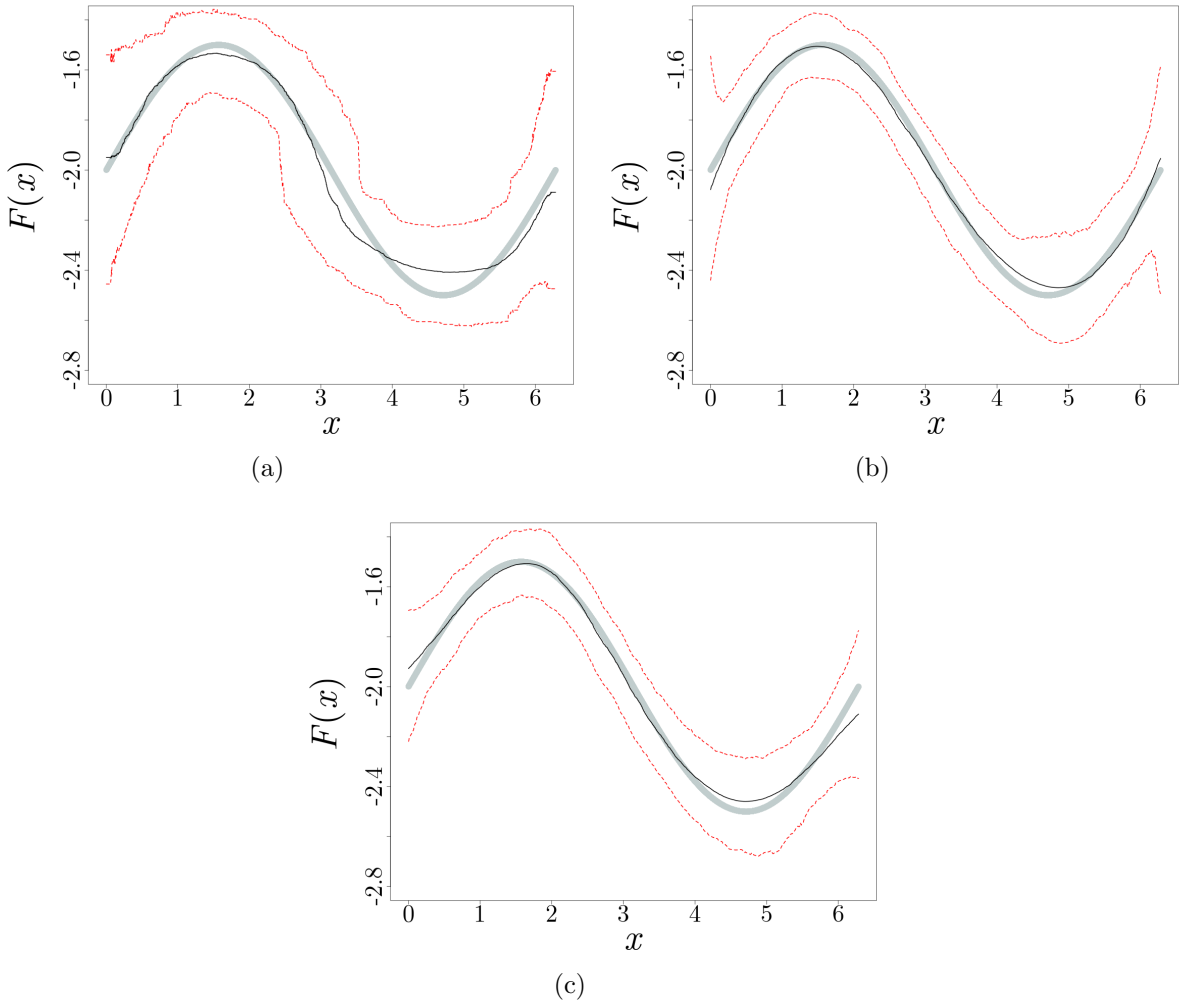

Figure 1: Simulation results for group-testing data generated from the predictor relationship in Equation (5). Gradient boosting with (a) regression trees, (b) splines, and (c) kernel smoothing as the weak learner was used to obtain model fits for each of 500 simulated data sets. The solid gray curve represents the true predictor function used to generate individual disease statuses. The solid black lines represent the average model fit, and the dashed red lines indicate the 0.025 and 0.975 quantiles for the 500 model fits.

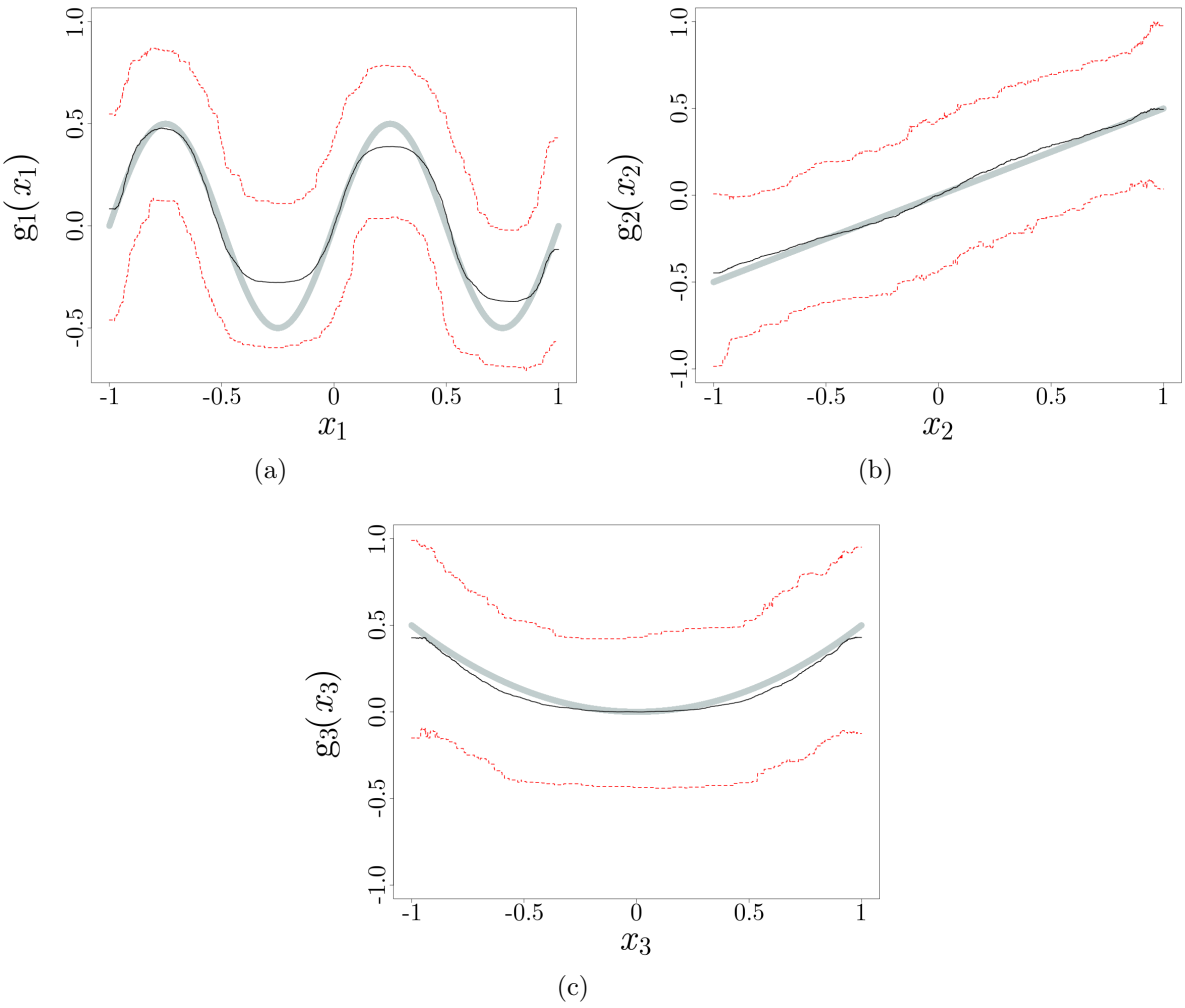

Figure 2: Simulation results for each of the functions in Equation (7). Gradient boosting with regression trees as the weak learner was used to obtain model fits for each of 500 simulated data sets. The solid gray curve represents (a) $g_1(\cdot)$, (b) $g_2(\cdot)$, and (c) $g_3(\cdot)$. The solid black lines represent the pointwise average fit for each function. We estimated the intercept $\widehat{\beta}_0$ by evaluating the model fit with all predictors at 0, i.e. $\widehat{F}(\mathbf{0}, \mathbf{0}, \mathbf{0})$. Then each function $\widehat{g}_j$ was estimated by evaluating one predictor variable across a sequence from -1 to 1 with the other two set to 0, and subtracting off $\widehat{\beta}_0$. The dashed red lines indicate the 0.025 and 0.975 quantiles for the 500 estimates (a) $\widehat{g}_1(\cdot)$, (b) $\widehat{g}_2(\cdot)$, and (c) $\widehat{g}_3(\cdot)$.

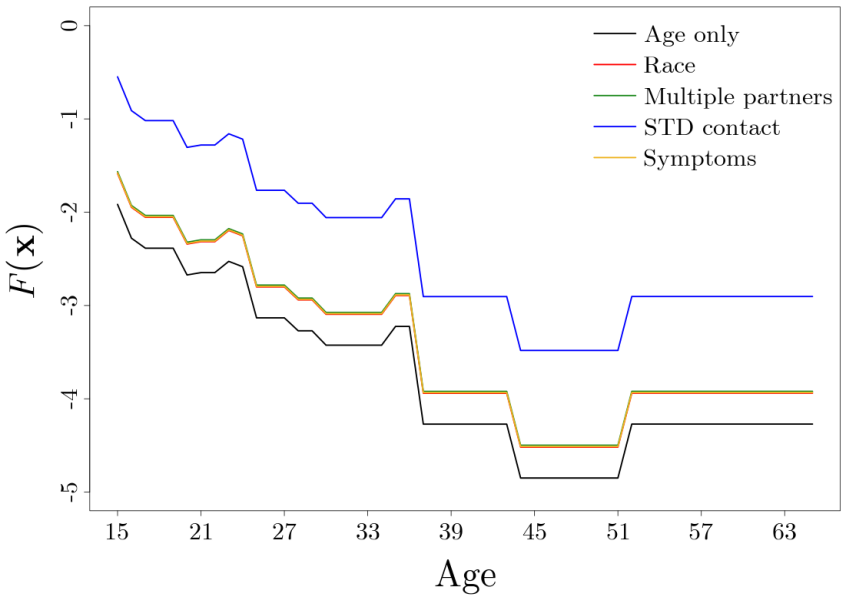

Figure 3: The black line plots the estimate of $F(\cdot)$ across the age predictor when all binary predictors are set to 0. The 4 remaining lines indicate the estimate when one binary predictor is set to 1 at a time, while age is also in the model. The 3 overlapping lines in the middle of the plot indicate that the estimated infection probabilities from being non-Caucasian, reporting multiple sexual partners, and exhibiting symptoms of infection have very similar values when age is the only other non-zero predictor in the model.

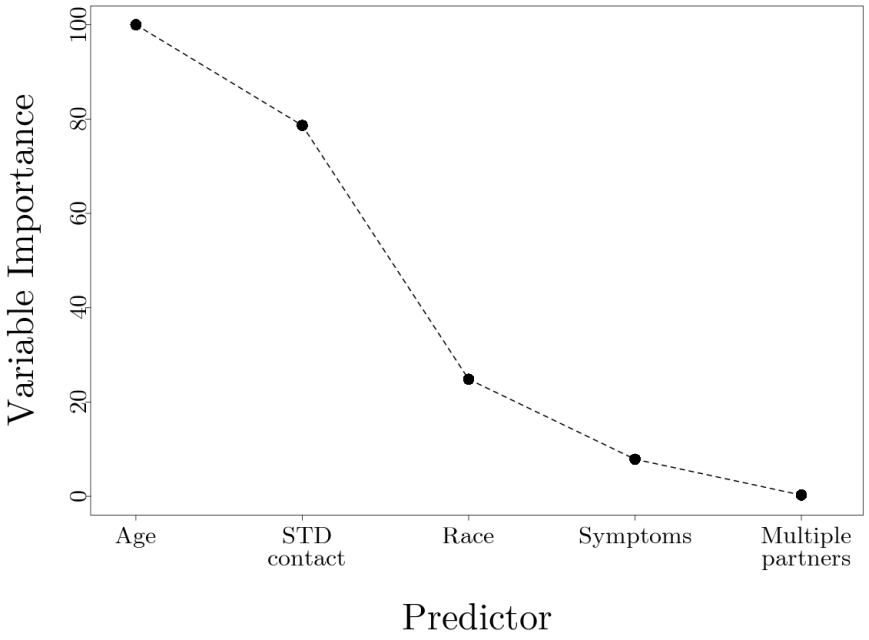

Figure 4: Variable importance measures for all 5 predictors for the SHL chlamydia data, scaled between 0 and 100.

Predictor	**Increase in log odds**
Race	0.331
Multiple partners	0.351
STD contact	1.367
Symptoms	0.338

Table 1: Estimated increase in the log odds of chlamydia infection from each indicator variable for the SHL chlamydia data.