# Web-based Supplementary Materials for "Bayesian Additive Regression Trees for Group Testing Data"

## A BART backfitting algorithm details

In this section, we describe the details of the Bayesian backfitting MCMC algorithm, outlined in Chipman et al. (2010), to sample from the posterior distribution of the regression trees $(T_1, M_1), (T_2, M_2), \ldots, (T_K, M_K)$. The algorithm iteratively fits the $k$th tree using the residuals based on a fit excluding the $k$th tree. In general, the algorithm is simply a Gibbs sampler that employs a modified version of Bayesian backfitting MCMC introduced by Hastie and Tibshirani (2000).

An iteration of the backfitting algorithm first requires $N$ successive draws of the latent random variables $\omega_i$ that were introduced in the second stage of our data augmentation procedure:

$$\omega_i \sim \begin{cases} TN\left[\eta_i, 1, (0, \infty)\right], & \text{if } \widetilde{Y}_i = 1 \\ TN\left[\eta_i, 1, (-\infty, 0)\right], & \text{if } \widetilde{Y}_i = 0, \end{cases}$$

where $TN\left[\mu, \sigma^2, (a, b)\right]$ denotes a truncated normal distribution with mean $\mu$ and variance $\sigma^2$, and support over the interval $(a, b)$. Then, we can treat the latent variables $\omega_i$ as continuous outcomes and recast our BART model as

$$\omega_i = \eta(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

for $i = 1, \ldots, N$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$ because we've employed the probit link. Following this, the algorithm then requires $K$ successive draws of the individual trees $(T_k, M_k)$ conditioning on the remaining $K - 1$ trees:

$$\pi\left((T_k, M_k) \,\middle|\, \boldsymbol{T}_{-\boldsymbol{k}}, \boldsymbol{M}_{-\boldsymbol{k}}, \boldsymbol{\omega}\right), \tag{2}$$

where $\boldsymbol{T}_{-\boldsymbol{k}}$ is the set of $K-1$ tree structures excluding $T_k$, and $\boldsymbol{M}_{-\boldsymbol{k}}$ are the associated terminal node parameters. To obtain a draw from (2), note that $\pi\left((T_k, M_k) \,\middle|\, \boldsymbol{T}_{-\boldsymbol{k}}, \boldsymbol{M}_{-\boldsymbol{k}}, \boldsymbol{\omega}\right)$ depends

on $(\boldsymbol{T_{-k}}, \boldsymbol{M_{-k}}, \boldsymbol{\omega})$ through the $k$th vector set of partial residuals $\boldsymbol{R_k} = (R_{k1}, \ldots, R_{kN})'$, where the $i$th element of $\boldsymbol{R_k}$ is given by

$$R_{ki} = \omega_i - \sum_{u \neq k}^{K} g(\boldsymbol{x}_i; T_u, M_u),$$

for $i = 1, \ldots, N$. Thus, our recasted model (1) can be temporarily reparameterized in terms of these partial residuals. That is,

$$R_{ki} \sim N\left(g(\boldsymbol{x}_i; T_k, M_k), 1\right),$$

and a posterior draw from (2) is equivalent to a posterior draw from a single regression tree $R_{ki} = g(\boldsymbol{x}_i; T_k, M_k) + \epsilon_i$; i.e., a posterior draw from

$$\pi\left((T_k, M_k) \,\middle|\, \boldsymbol{R_k}\right). \tag{3}$$

We can obtain a draw from (3) in two successive steps. Since a conjugate normal prior on $\mu_{kt}$ was employed, for $t = 1, \ldots, b_k$, we can first integrate out $M_k$ and sample from $\pi\left(T_k \mid \boldsymbol{R_k}\right)$. Then, we can obtain a draw from $\pi\left(M_k \mid T_k, \boldsymbol{R_k}\right)$.

We obtain a draw from $\pi\left(T_k \mid \boldsymbol{R_k}\right)$ using the Metropolis-Hastings (MH) algorithm of Chipman et al. (1998), where we first generate a candidate tree $T_k^*$ with probability distribution $q(T_k, T_k^*)$ and accept $T_k^*$ with probability

$$\alpha(T_k, T_k^*) = \min\left\{1, \frac{q(T_k^*, T_k)}{q(T_k, T_k^*)} \frac{p(\boldsymbol{R_k} \mid T_k^*, M_k)}{p(\boldsymbol{R_k} \mid T_k, M_k)} \frac{\pi(T_k^*)}{\pi(T_k)}\right\}, \tag{4}$$

where $\frac{q(T_k^*, T_k)}{q(T_k, T_k^*)}$ is the transition ratio, $\frac{p(\boldsymbol{R_k}|T_k^*, M_k)}{p(\boldsymbol{R_k}|T_k, M_k)}$ is the likelihood ratio, and $\frac{\pi(T_k^*)}{\pi(T_k)}$ is the tree structure ratio. A new tree $T_k^*$ can be proposed given the current tree $T_k$ using one of four moves: growing a terminal node; pruning a pair of terminal nodes; swapping the splitting criteria of two non-terminal nodes; and changing the splitting criteria of a non-terminal node. For further details, see Chipman et al. (1998, 2010).

Once we have the draw from $\pi\left(T_k \mid \boldsymbol{R_k}\right)$, the posterior draw from $\pi\left(M_k \mid T_k, \boldsymbol{R_k}\right)$ is a set of independent draws of the terminal node parameters $\mu_{kt}$ from a normal distribution. Refer to Web Appendix B for its derivation and complete expression.

## B Posterior distribution for $\mu_{kt}$

Let $\boldsymbol{R_{k(t)}}$ be the $n_t$-dimensional subset vector of $\boldsymbol{R_k}$, where $n_t$ is the number of elements of $\boldsymbol{R_k}$ allocated to the terminal node parameter $\mu_{kt}$. Note that $R_{k(t)h} \,\big|\, T_k, M_k \sim N(\mu_{kt}, 1)$, for $h = 1, \ldots, n_t$, and $\mu_{kt} \mid T_k \sim N(0, \sigma_\mu^2)$. Then, we derive the posterior distribution of $\mu_{kt}$ as follows:

$$\begin{aligned}
\pi(\mu_{kt} \mid T_k, \boldsymbol{R_k}) &\propto \pi\left(\boldsymbol{R_{k(t)}} \,\middle|\, T_k, \mu_{kt}\right) \pi(\mu_{kt} \mid T_k) \\
&\propto \exp\left\{-\frac{\sum_h (R_{k(t)h} - \mu_{kt})^2}{2}\right\} \exp\left\{-\frac{\mu_{kt}^2}{2\sigma_\mu^2}\right\} \\
&\propto \exp\left\{-\frac{(n_t\sigma_\mu^2 + 1)\mu_{kt}^2 - 2(\sigma_\mu^2 \sum_h R_{k(t)h})\mu_{kt}}{2\sigma_\mu^2}\right\} \\
&\propto \exp\left\{-\frac{\left(\mu_{kt} - \frac{\sigma_\mu^2 \sum_h R_{k(t)h}}{n_t\sigma_\mu^2+1}\right)^2}{2\frac{\sigma_\mu^2}{n_t\sigma_\mu^2+1}}\right\}.
\end{aligned}$$

Therefore, the posterior distribution of $\mu_{kt}$ is given by

$$\mu_{kt} \mid T_k, \boldsymbol{R_k} \sim N\left(\frac{\sigma_\mu^2 \sum_h R_{k(t)h}}{n_t\sigma_\mu^2 + 1},\ \frac{\sigma_\mu^2}{n_t\sigma_\mu^2 + 1}\right).$$

## C Posterior sampling algorithm

1. Initialize $(T_1^{(0)}, M_1^{(0)}), \ldots, (T_K^{(0)}, M_K^{(0)})$ and $\widetilde{Y}_i^{(0)}$ for $i = 1, \ldots, N$. If estimating assay accuracies, then also initialize $\boldsymbol{S_e}^{(0)}$ and $\boldsymbol{S_p}^{(0)}$. Set $s = 1$.
   If not estimating assay accuracies, set $\boldsymbol{S_e}^{(s)} = \boldsymbol{S_e}$ and $\boldsymbol{S_p}^{(s)} = \boldsymbol{S_p}$ for all $s$.

2. For $i = 1, \ldots, N$, sample

$$\omega_i^{(s)} \sim \begin{cases} TN\left[\eta_i, 1, (0, \infty)\right], & \text{if } \widetilde{Y}_i^{(s-1)} = 1 \\ TN\left[\eta_i, 1, (-\infty, 0)\right], & \text{if } \widetilde{Y}_i^{(s-1)} = 0, \end{cases}$$

   where $\eta_i = \eta(\boldsymbol{x}_i) = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k^{(s-1)}, M_k^{(s-1)})$. Aggregate $\boldsymbol{\omega}^{(s)} = (\omega_1^{(s)}, \ldots, \omega_N^{(s)})'$.

3. For $k = 1, \ldots, K$, sample $(T_k^{(s)}, M_k^{(s)})$ from $\pi\left((T_k, M_k) \mid (\boldsymbol{T}_{-\boldsymbol{k}}^{(s)}, \boldsymbol{M}_{-\boldsymbol{k}}^{(s)}), \boldsymbol{\omega}^{(s)}\right)$, where

$$(\boldsymbol{T}_{-\boldsymbol{k}}^{(s)}, \boldsymbol{M}_{-\boldsymbol{k}}^{(s)}) = \left((T_1^{(s)}, M_1^{(s)}), \ldots, (T_{k-1}^{(s)}, M_{k-1}^{(s)}), (T_{k+1}^{(s-1)}, M_{k+1}^{(s-1)}), \ldots, (T_K^{(s-1)}, M_K^{(s-1)})\right)',$$

   to obtain $\eta_i = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k^{(s)}, M_k^{(s)})$ for $i = 1, \ldots, N$.

4. If estimating assay accuracy probabilities, then for $l = 1, \ldots, L$, sample $S_{e(l)}^{(s)} \sim Beta(a_{e(l)}^*, b_{e(l)}^*)$ and $S_{p(l)}^{(s)} \sim Beta(a_{p(l)}^*, b_{p(l)}^*)$, where $a_{e(l)}^*$, $b_{e(l)}^*$, $a_{p(l)}^*$, and $b_{p(l)}^*$ are evaluated at $\widetilde{\boldsymbol{Y}}^{(s-1)}$. Aggregate $\boldsymbol{S_e}^{(s)} = \left(S_{e(1)}^{(s)}, \ldots, S_{e(L)}^{(s)}\right)'$ and $\boldsymbol{S_p}^{(s)} = \left(S_{p(1)}^{(s)}, \ldots, S_{p(L)}^{(s)}\right)'$.

5. For $i = 1, \ldots, N$, sample

$$\widetilde{Y}_i^{(s)} \sim Bernoulli\left(\frac{p_{i1}^*}{p_{i0}^* + p_{i1}^*}\right),$$

   where $p_{i0}^*$ and $p_{i1}^*$ are evaluated at $\widetilde{\boldsymbol{Y}}_{-\boldsymbol{i}}^{(s)} = \left(\widetilde{Y}_1^{(s)}, \ldots, \widetilde{Y}_{i-1}^{(s)}, \widetilde{Y}_{i+1}^{(s-1)}, \ldots, \widetilde{Y}_N^{(s-1)}\right)'$, $\boldsymbol{S_e}^{(s)}$, $\boldsymbol{S_p}^{(s)}$, and $\eta_i = \sum_{k=1}^{K} g(\boldsymbol{x}_i; T_k^{(s)}, M_k^{(s)})$.

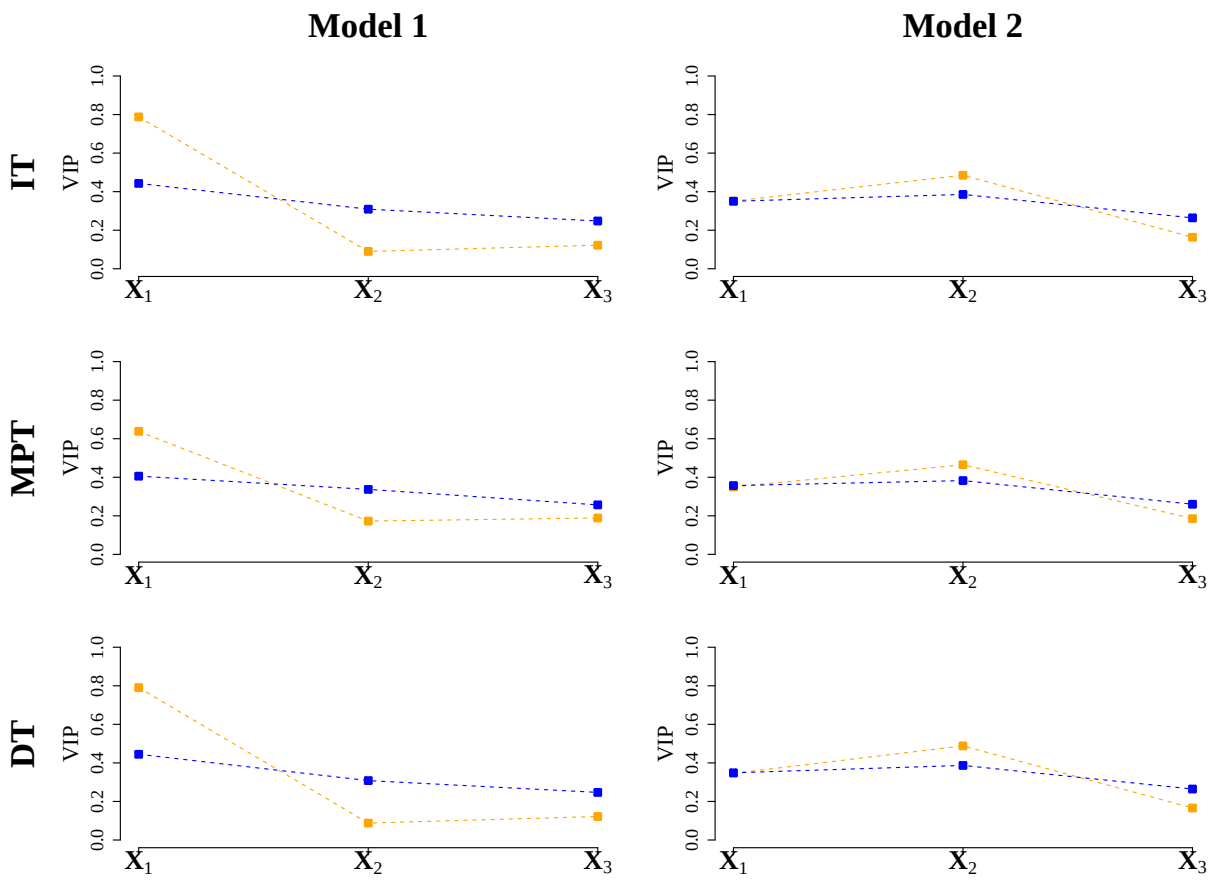
6. Increment $s = s + 1$ and return to Step 2.

## D Additional simulation results

This section provides additional simulation results from the numerical studies of Section 4.

Web Table 1: Simulation results for the three model configurations when assay accuracy probabilities are **known**: Average estimated AUC and sample standard deviation (in parentheses).

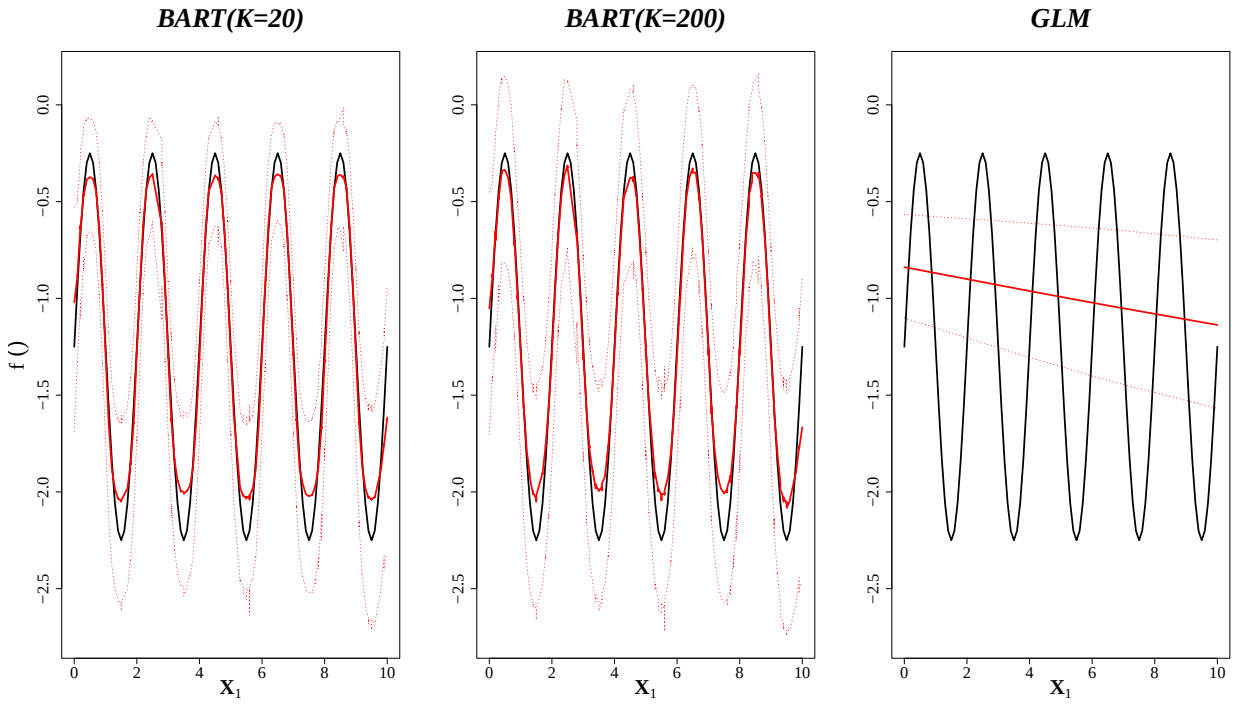
| Model | GT Protocol | | BART($K$=20) | BART($K$=200) | GLM |
|---|---|---|---|---|---|
| M1 | IT | In-Sample | 0.80 (0.01) | 0.81 (0.01) | 0.54 (0.01) |
| | | Out-of-Sample | 0.77 (0.02) | 0.78 (0.02) | 0.53 (0.02) |
| | MPT | In-Sample | 0.76 (0.01) | 0.77 (0.01) | 0.54 (0.01) |
| | | Out-of-Sample | 0.74 (0.02) | 0.75 (0.02) | 0.52 (0.02) |
| | DT | In-Sample | 0.80 (0.01) | 0.82 (0.01) | 0.54 (0.01) |
| | | Out-of-Sample | 0.77 (0.02) | 0.78 (0.02) | 0.53 (0.02) |
| M2 | IT | In-Sample | 0.99 (0.00) | 0.99 (0.00) | 0.99 (0.00) |
| | | Out-of-Sample | 0.98 (0.00) | 0.98 (0.00) | 0.98 (0.00) |
| | MPT | In-Sample | 0.98 (0.0) | 0.99 (0.0) | 0.98 (0.0) |
| | | Out-of-Sample | 0.97 (0.0) | 0.98 (0.0) | 0.98 (0.0) |
| | DT | In-Sample | 0.99 (0.0) | 0.99 (0.0) | 0.98 (0.0) |
| | | Out-of-Sample | 0.98 (0.0) | 0.98 (0.0) | 0.98 (0.0) |



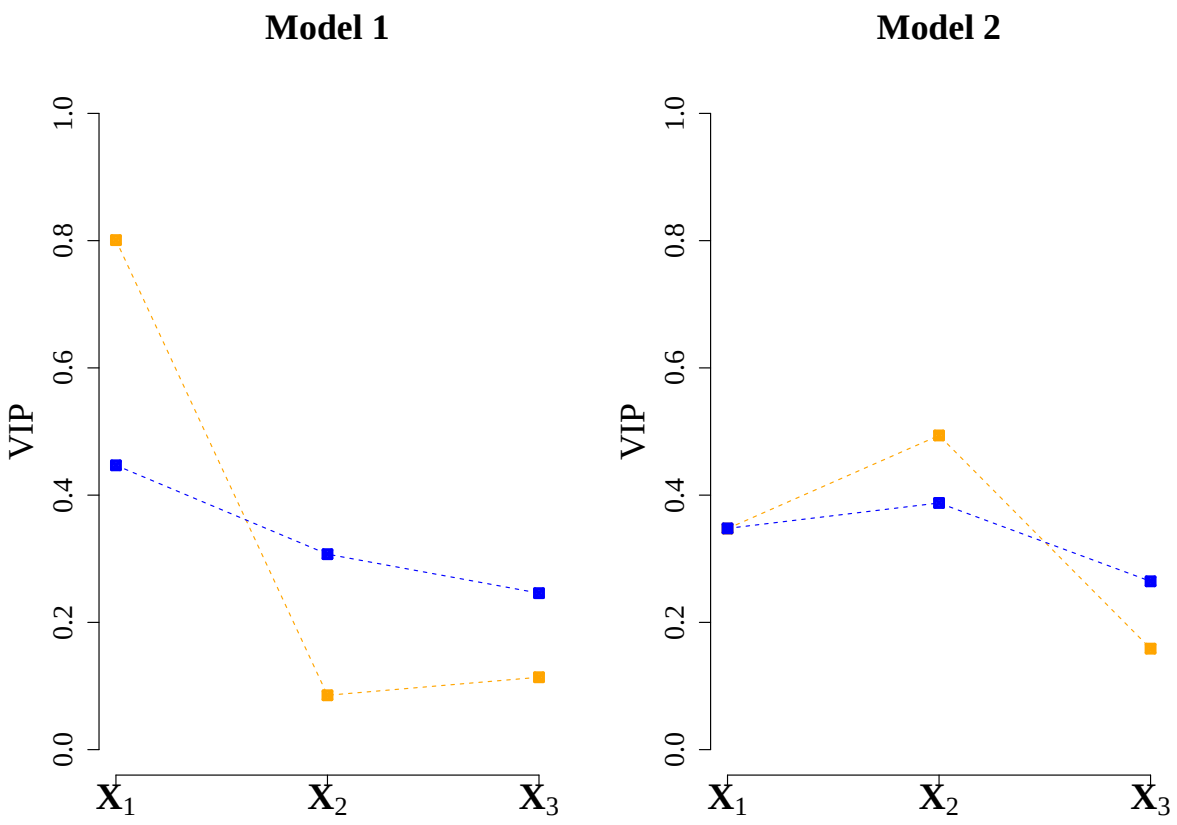

Web Figure 1: Simulation results for population models M1 (left column) and M2 (right column) when assay accuracy probabilities are **known**: Variable inclusion proportions (VIPs), averaged over the 500 simulations, for BART with $K$=20 trees (orange) and $K$=200 trees (blue) under the IT (top row), MPT (middle row), and DT (bottom row) protocols.

Web Table 2: Simulation results for the three model configurations when assay accuracy probabilities are **unknown**: Average estimated AUC scores (and sample standard deviation in parentheses).

| Model | | BART($K$=20) | BART($K$=200) | GLM |
|---|---|---|---|---|
| M1 | In-Sample | 0.80 (0.01) | 0.82 (0.01) | 0.54 (0.01) |
| | Out-of-Sample | 0.77 (0.02) | 0.78 (0.02) | 0.53 (0.02) |
| M2 | In-Sample | 0.99 (0.00) | 0.99 (0.00) | 0.99 (0.00) |
| | Out-of-Sample | 0.98 (0.00) | 0.98 (0.00) | 0.98 (0.00) |




Web Figure 2: In-sample simulation results for the three model configurations when assay accuracy probabilities are **unknown**: BART $K$=20 (left), BART $K$=200 (middle), and GLM (right) under the DT protocol. 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).




Web Figure 3: Simulation results for population models M1 (left) and M2 (right) when assay accuracy probabilities are **unknown**: Variable inclusion proportions (VIPs), averaged over the 500 simulations, for BART with $K$=20 trees (orange) and $K$=200 trees (blue).

Web Table 3: Simulation results for population models M1 and M2 under DT protocol when assay accuracy probabilities are **unknown**: Average bias of the 500 posterior mean estimates (Bias), sample standard deviation of the 500 posterior mean estimates (SSD), average of the 500 estimated of the posterior standard deviation (ESE), and empirical coverage probability (CP95) of nominal 95% equal-tail credible intervals are reported for each parameter.

| Model | | $S_{e(1)} = 0.95$ | $S_{p(1)} = 0.98$ | $S_{e(2)} = 0.98$ | $S_{p(2)} = 0.99$ |
|---|---|---|---|---|---|
| M1/BART($K$=20) | Bias (CP95) | -0.02 (0.96) | 0.00 (0.99) | 0.00 (0.99) | 0.00 (0.98) |
| | SSD (ESE) | 0.03 (0.04) | 0.01 (0.01) | 0.01 (0.01) | 0.01 (0.01) |
| M1/BART($K$=200) | Bias (CP95) | -0.02 (0.97) | 0.00 (1.00) | 0.00 (0.99) | 0.00 (1.00) |
| | SSD (ESE) | 0.03 (0.04) | 0.01 (0.01) | 0.01 (0.01) | 0.00 (0.01) |
| M1/GLM | Bias (CP95) | -0.03 (1.00) | 0.00 (1.00) | -0.01 (1.00) | 0.00 (1.00) |
| | SSD (ESE) | 0.02 (0.05) | 0.00 (0.01) | 0.01 (0.02) | 0.00 (0.01) |
| M2/BART($K$=20) | Bias (CP95) | -0.01 (0.93) | 0.00 (0.95) | 0.00 (0.93) | 0.00 (0.95) |
| | SSD (ESE) | 0.01 (0.01) | 0.01 (0.01) | 0.01 (0.01) | 0.00 (0.00) |
| M2/BART($K$=200) | Bias (CP95) | -0.01 (0.86) | 0.00 (0.96) | 0.00 (0.92) | 0.00 (0.94) |
| | SSD (ESE) | 0.01 (0.01) | 0.01 (0.01) | 0.01 (0.01) | 0.00 (0.00) |
| M2/GLM | Bias (CP95) | 0.00 (0.95) | 0.00 (0.97) | 0.00 (0.95) | 0.00 (0.95) |
| | SSD (ESE) | 0.01 (0.01) | 0.01 (0.01) | 0.01 (0.01) | 0.00 (0.00) |

# E Aptima Combo 2 Assay (AC2A) accuracy

Here we summarize the AC2A accuracy based on data from a pilot study and describe how to incorporate this information into our model used for the data application in Section 5, when assay accuracy is unknown and to be estimated. This section is modified from Web Appendix D of McMahan et al. (2017).

Web Table 4: AC2A pilot data.

| Stratum | TP | FN | TN | FP | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Female/Swab | 195 | 12 | 1154 | 28 | $S_{e(1)} = 0.942$ | $S_{p(1)} = 0.976$ |
| Female/Urine | 197 | 11 | 1170 | 13 | $S_{e(2)} = 0.947$ | $S_{p(2)} = 0.989$ |
| Male/Swab | 260 | 11 | 774 | 20 | $S_{e(3)} = 0.959$ | $S_{p(3)} = 0.975$ |
| Male/Urine | 276 | 6 | 801 | 12 | $S_{e(4)} = 0.979$ | $S_{p(4)} = 0.985$ |

The notation used in Web Table 4 is defined below.

$$\begin{aligned} \text{TP} &= \text{ number of true positive individual test results} \\ \text{FN} &= \text{ number of false negative individual test results} \\ \text{TN} &= \text{ number of true negative individual test results} \\ \text{FP} &= \text{ number of false positive individual test results} \end{aligned}$$

Recall from Section 2.1 that we place independent Beta priors on the assay accuracies, chosen for computational convenience. To incorporate our prior belief about the assay sensitivity and specificity based on the pilot data, we create informative Beta priors by specifying the hyperparameter values as follows:

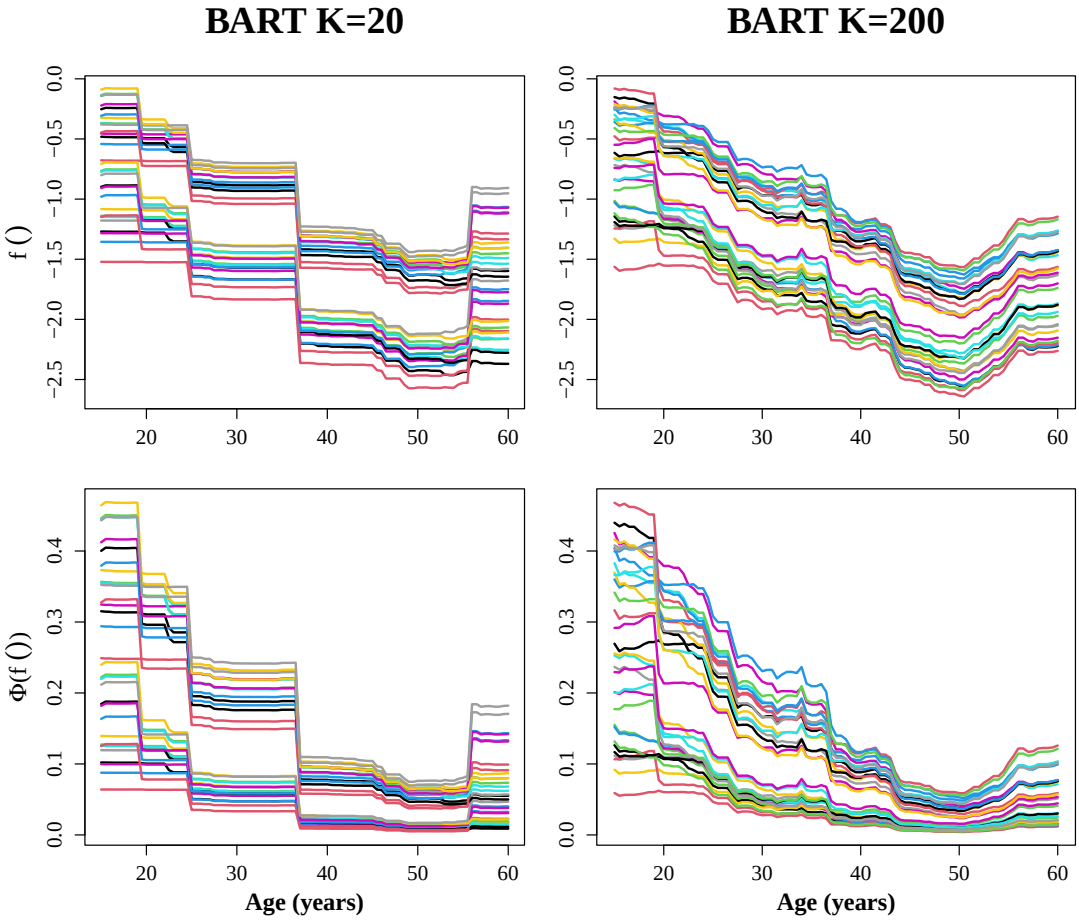
$$S_e \sim Beta(\text{TP} + 1, \text{FN} + 1)$$

$$S_p \sim Beta(\text{TN} + 1, \text{FP} + 1).$$

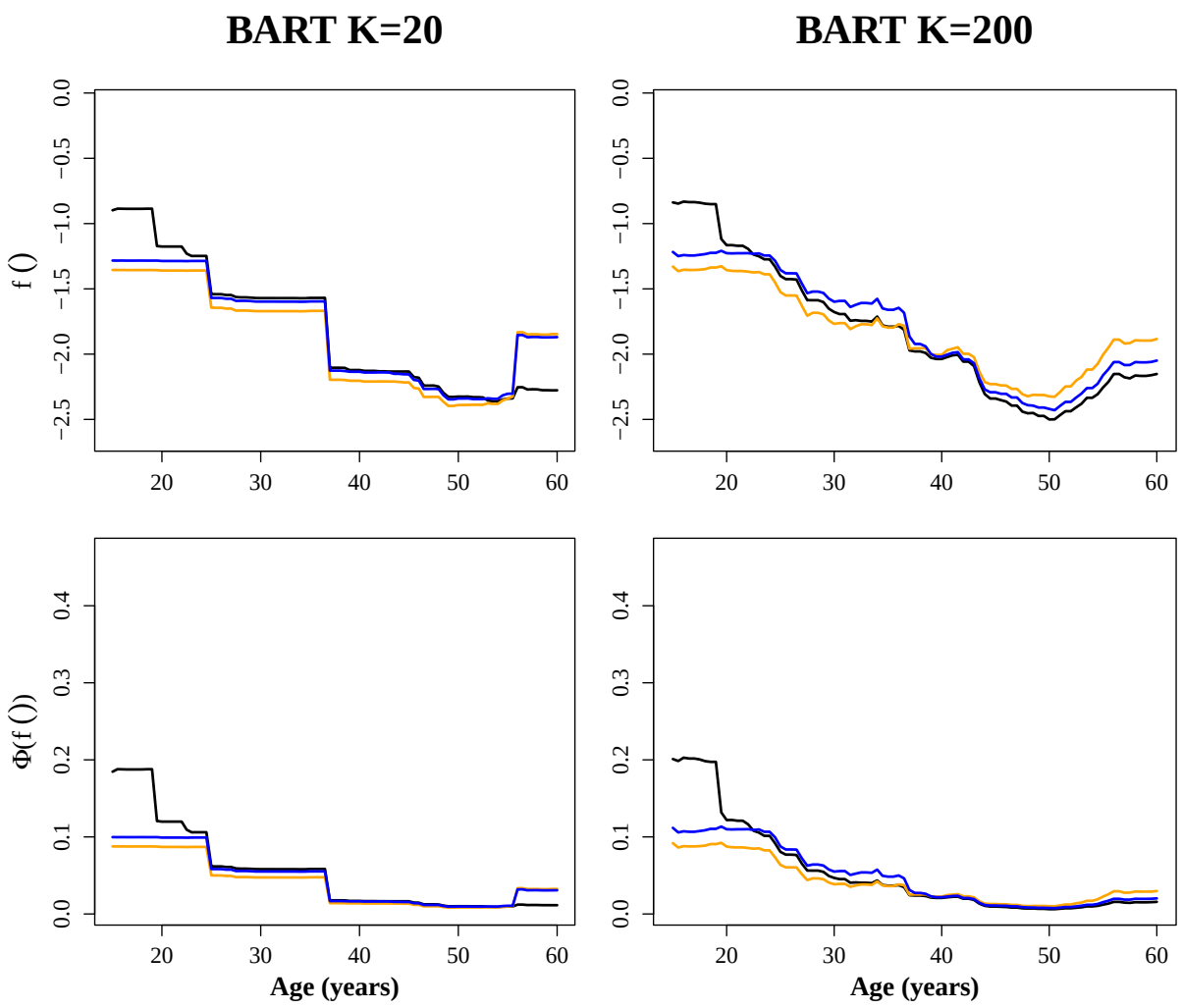
With this, the prior distributions for $S_e$ and $S_p$ are concentrated around $\text{TP}/(\text{TP} + \text{FN})$ and TN/(TN+FP), respectively. In particular, for swab specimens, we specify $S_e \sim Beta(196, 123)$ and $S_p \sim Beta(1156, 29)$; for urine specimens, we specify $S_e \sim Beta(198, 12)$ and $S_p \sim Beta(1171, 13)$.

# F Additional Iowa chlamydia data analysis results

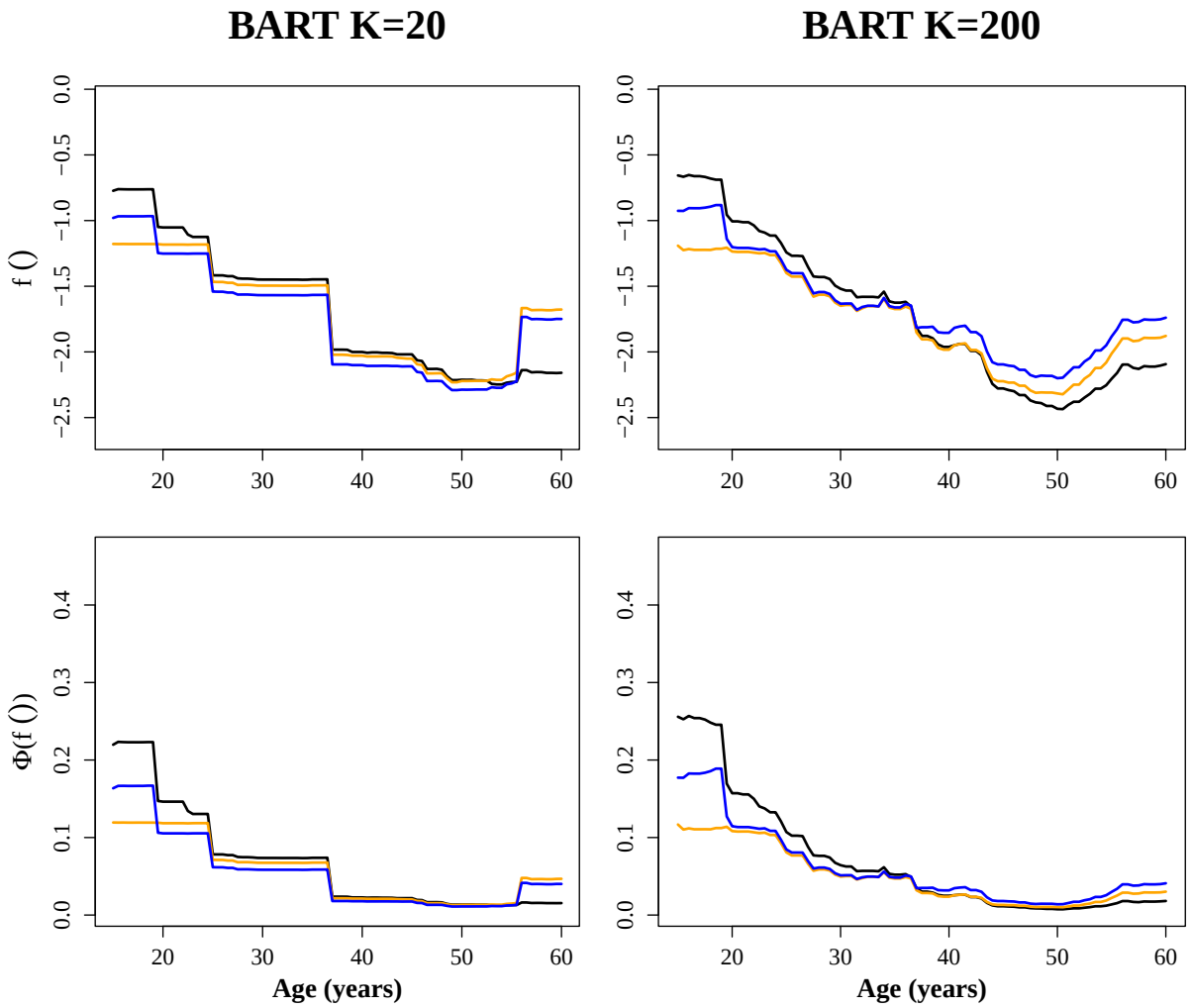
This section provides additional estimation results from the Iowa chlamydia data analysis of Section 5.



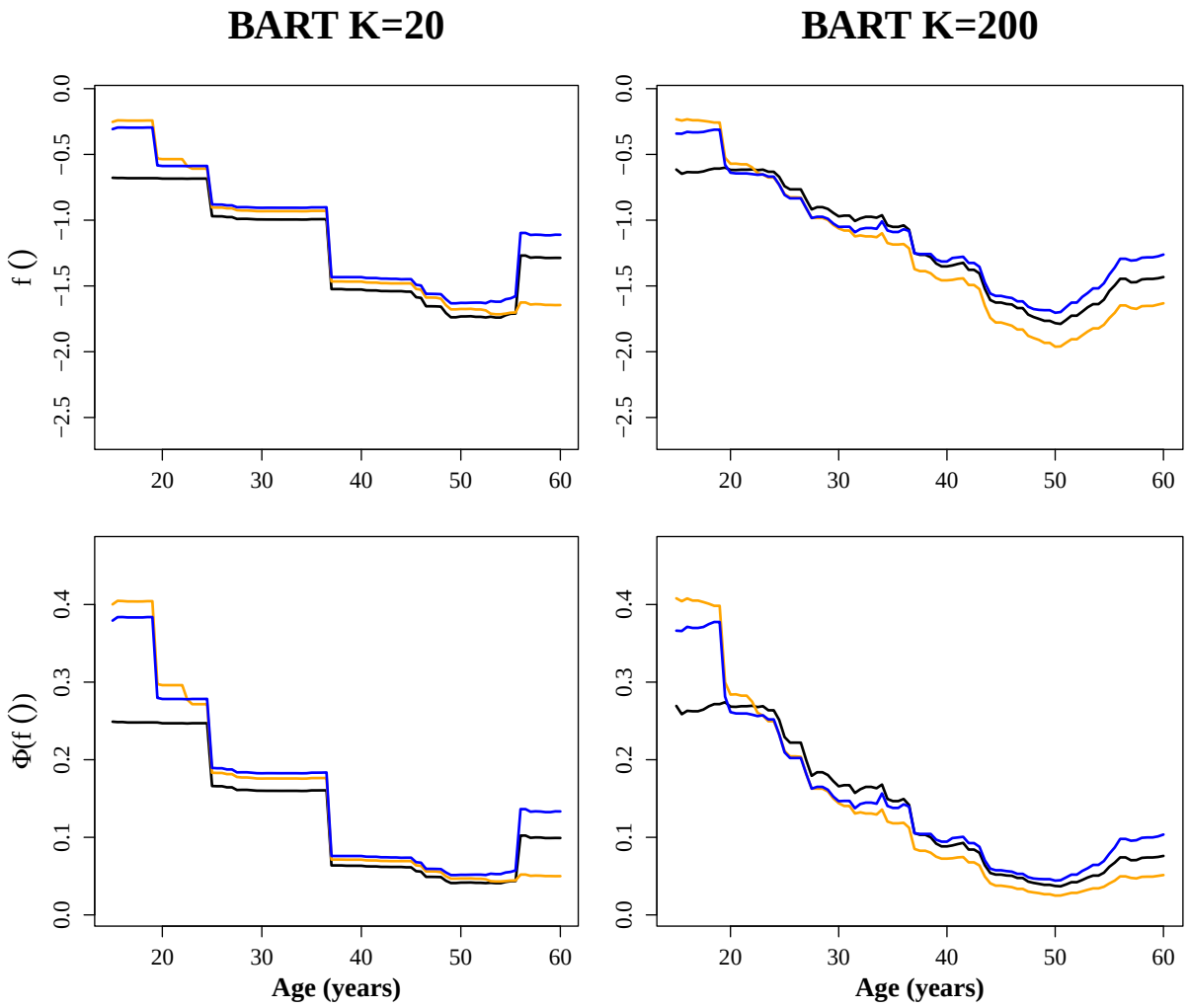

Web Figure 4: Iowa Chlamydia Data. Posterior mean estimates of the function $f(\cdot)$ (top row) and the probabilities $\Phi(f(\cdot))$ (bottom row) from the BART configurations with $K$=20 trees (left) and $K$=200 trees (right), plotted against the age covariate $x_{i1}$ for each of the risk profiles (i.e., for all 32 combinations of the 5 binary risk factors).



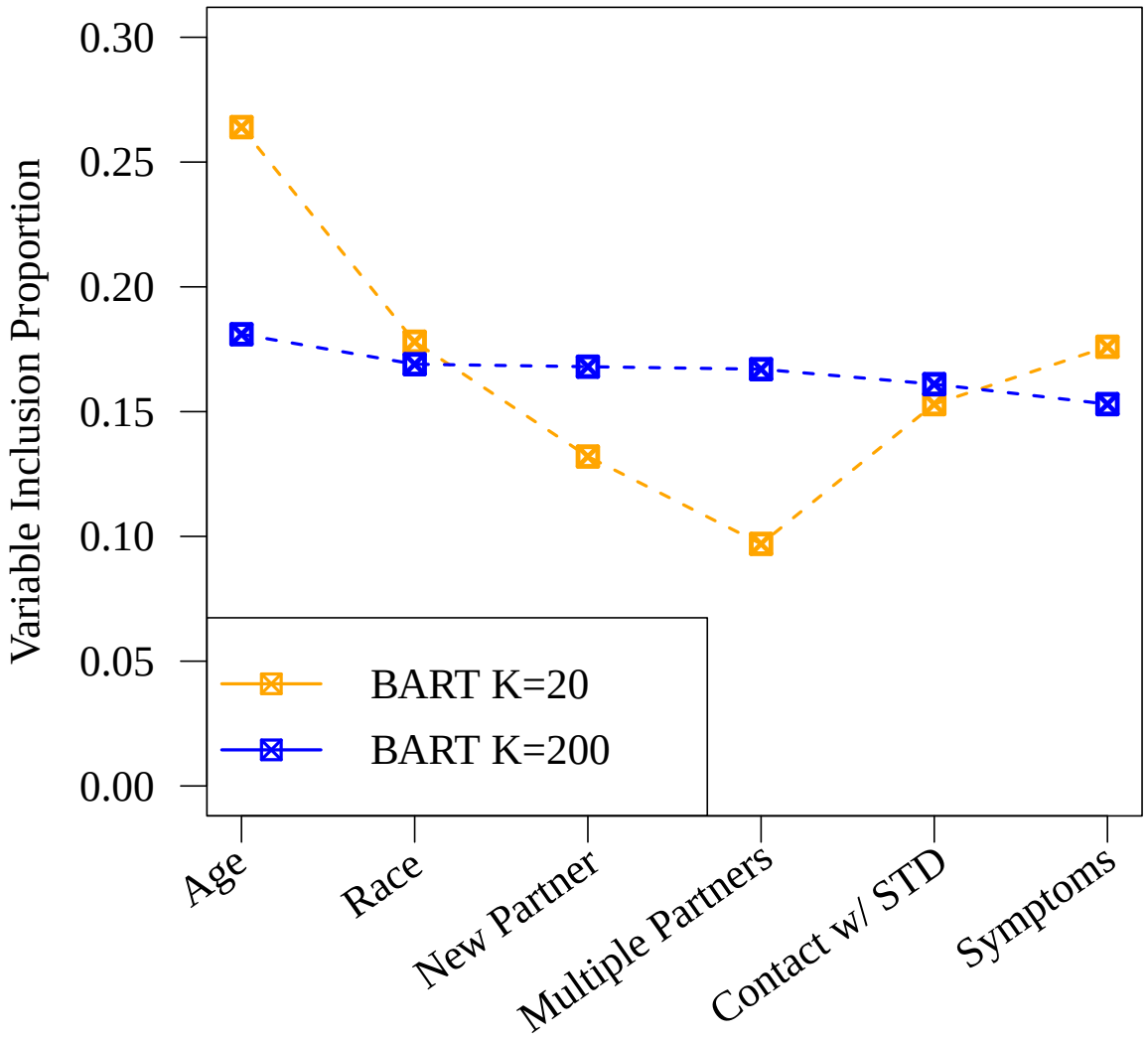

Web Figure 5: Iowa Chlamydia Data. Posterior mean estimates of the function $f(\cdot)$ (top row) and the probabilities $\Phi\left(f(\cdot)\right)$ (bottom row) from the BART configurations with $K$=20 trees (left) and $K$=200 trees (right), plotted against the age covariate $x_{i1}$ for three risk profiles: non-Caucasian patients that presented symptoms of infection (black curve); Caucasian patients that reported a new sexual partner (orange curve); Caucasian patients that reported having multiple sexual partners (blue curve).




Web Figure 6: Iowa Chlamydia Data. Posterior mean estimates of the function $f(\cdot)$ (top row) and 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 having multiple sexual partners and presented symptoms of infection (black curve); Caucasian patients that reported having a new sexual partner and multiple sexual partners (orange curve); Caucasian patients that reported having a new sexual partner and presented symptoms of infection (blue curve).




Web Figure 7: Iowa Chlamydia Data. Posterior mean estimates of the function $f(\cdot)$ (top row) and 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: Caucasian patients that reported sexual contact with an STD-positive partner (black curve); non-Caucasian patients that reported sexual contact with an STD-positive partner and presented symptoms of infection (orange curve); Caucasian patients that reported a new sexual partner, sexual contact with an STD-positive partner, and presented symptoms of infection (blue curve).




Web Figure 8: Iowa Chlamydia Data. Variable inclusion proportions for the BART models with 20 (orange) and 200 (blue) trees.

Web Table 5: Iowa Chlamydia Data. Results from estimating the assay accuracy probabilities $S_{e(l)}$ and $S_{p(l)}$, for $l = 1, 2, 3$. Posterior mean estimates (Est), estimated posterior standard deviations (ESE), and 95% equal-tail credible intervals (CI95) are provided.

| | | **BART($K$=20)** | | | **BART($K$=200)** | | | **GLM** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Param. | Descrip. | Est | ESE | CI95 | Est | ESE | CI95 | Est | ESE | CI95 |
| $S_{e(1)}$ | Swab Ind. | 0.98 | 0.00 | (0.97, 0.99) | 0.98 | 0.00 | (0.97, 0.99) | 0.97 | 0.04 | (0.83, 0.99) |
| $S_{e(2)}$ | Urine Ind. | 0.95 | 0.02 | (0.91, 0.97) | 0.95 | 0.02 | (0.91, 0.97) | 0.90 | 0.10 | (0.56, 0.96) |
| $S_{e(3)}$ | Swab Pool | 0.94 | 0.02 | (0.91, 0.97) | 0.94 | 0.02 | (0.91, 0.97) | 0.90 | 0.10 | (0.57, 0.97) |
| $S_{p(1)}$ | Swab Ind. | 0.97 | 0.00 | (0.97, 0.98) | 0.97 | 0.00 | (0.97, 0.98) | 0.97 | 0.00 | (0.96, 0.98) |
| $S_{p(2)}$ | Urine Ind. | 0.99 | 0.00 | (0.98, 0.99) | 0.99 | 0.00 | (0.98, 0.99) | 0.99 | 0.00 | (0.98, 0.99) |
| $S_{p(3)}$ | Swab Pool | 0.99 | 0.00 | (0.99, 0.99) | 0.99 | 0.00 | (0.99, 0.99) | 0.99 | 0.00 | (0.98, 0.99) |

## References

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

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93:935–948, 1998.

T. Hastie and R. Tibshirani. Bayesian backfitting. *Statistical Science*, 15:196–213, 2000.

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