Strategies for variable selection in large-scale healthcare database studies with missing covariate and outcome data

Jung-Yi Joyce Lin 1, Liangyuan Hu∗,2, Chuyue Huang 3, Steven Lawrence 1, and Usha Govindarajulu 1

1 Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, New York 10029
2 Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, New Jersey 08854
3 Primary Research Solution LLC. 115 W 18th St, New York, New York 10011

Received zzz, revised zzz, accepted zzz

Prior work has shown that combining bootstrap imputation with tree-based machine learning variable selection methods can recover the good performance achievable on fully observed data when covariate and outcome data are missing at random (MAR). This approach however is computationally expensive, especially on large-scale datasets. We propose an inference-based method RR-BART, that leverages the likelihood-based Bayesian machine learning technique, Bayesian Additive Regression Trees, and uses Rubin’s rule to combine the estimates and variances of the variable importance measures on multiply imputed datasets for variable selection in the presence of missing data. A representative simulation study suggests that RR-BART performs at least as well as combining bootstrap with BART, BI-BART, but offers substantial computational savings, even in complex conditions of nonlinearity and nonadditivity with a large percentage of overall missingness under the MAR mechanism. RR-BART is also less sensitive to the end note prior via the hyperparameter k than BI-BART, and does not depend on the selection threshold value π as required by BI-BART. Our simulation studies also suggest that encoding the missing values of a binary predictor as a separate category significantly improves the power of selecting the binary predictor for BI-BART. We further demonstrate the methods via a case study of risk factors for 3-year incidence of metabolic syndrome with data from the Study of Women’s Health Across the Nation.

Key words: Missing at random; Multiply imputed datasets; Bootstrap imputation; Tree-based methods; Variable importance

1 Introduction

The problem of variable selection is one of the most popular model selection problems in statistical applications (George, 2000). Variable selection involves modeling the relationship between an outcome variable and a set of potential explanatory variables or predictors and identifying a subset of predictors that has the most impact on the model fit. There are a wide variety of approaches for variable selection in fully observed data. For example, stepwise methods such as the backward stepwise regression (Hocking, 1976) that sequentially deletes or adds variables based on hypothesis tests between nested models, and shrinkage methods such as the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) that optimizes a likelihood penalized for model complexity using certain criteria (e.g., the Akaike or Bayesian information criterion), have been widely used in practice. Heinze et al. (2018) provides a detailed review of the variable selection methods.

By comparison, variable selection in the presence of missing covariate and outcome data has been less studied. The challenge lies in the fact that there are different statistical approaches for handling missing data under different missing data mechanisms. As a result, methods for variable selection need to be

*Corresponding author: e-mail: liangyuan.hu@rutgers.edu, Phone: 732-235-4646, Fax: 732-235-5476
tailed to the specific statistical approach used under a given missing data mechanism. There are three general missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Little et al., 2012). When the missingness depends neither on observed data nor on the missing data, the data are said to be MCAR. Only when data are MCAR, the complete cases analysis would not bias the results as the complete cases are representative of the entire data. A more realistic assumption about the missing data is that the mechanism of missingness may depend on the observed data, and then the missing data are MAR given the observed data (Sterne et al., 2009). Under the MAR assumption, one can predict the missing values based on the observed data. In a more challenging situation where the missingness depends on the missing data, the data are MNAR (Little and Rubin, 2019). To handle MNAR, an approach recommended by the National Research Council (NATIONAL RESEARCH COUNCIL, 2010) is sensitivity analysis (Hogan et al., 2014; Hu et al., 2018) that evaluates the impact of the potential magnitude of departure from MAR on analysis results. Little and Rubin (2019) provide a comprehensive review of existing statistical approaches for handling missing data. In this paper, we focus on the MAR mechanism that allows replacing missing data with substituted values or imputation based on the observed data, which is widely accepted in epidemiological and health research.

Under MAR, variable selection has been conducted in combination with three commonly used statistical approaches for handling missing data: 1) inverse probability weighting and its variants; 2) likelihood-based methods; and 3) imputation for missing values (Long and Johnson, 2015). We refer to Long and Johnson (Long and Johnson, 2015) for an overview of these three approaches. In this article, we put the third approach, variable selection in conjunction with imputation, under close investigation. This approach is widely applicable and is amenable to general missing data patterns because once the missing values are imputed, a wide array of variable selection methods designed for complete data can be applied. The key consideration is the uncertainty about both the imputation and selecting a subset of important predictors.

Wood et al. (2008) compared strategies for combining the backward stepwise selection approach and multiple imputation for incomplete data, and recommended selecting variables based on Rubin’s rules that combine estimates of parameters and standard errors across multiply imputed data. Long and Johnson (2015) proposed combining bootstrap imputation and stability selection (Meinshausen and Bühlmann, 2010), and used simulations to demonstrate several advantages of this approach such as accommodating general missing data patterns and avoiding the issue of unstable weights associated with inverse probability weighting based methods. Stability selection is a general variable selection approach based on resampling or subsampling originally developed for the fully observed data (Meinshausen and Bühlmann, 2010). In stability selection, the randomized lasso is applied to each random sample, drawn by random subsampling or resampling of the complete data, and a subset of important predictors are selected based on the variable selection results across all $M$ random samples using a threshold $\pi$, e.g., if they are selected in at least $\pi M$ random samples. Both studies rely on parametric assumptions about the exact relationship between response and covariates, which may be misspecified and as a result, variable selection results can be sub-optimal (Bleich et al., 2014).

Flexible nonparametric methods can mitigate the reliance on the parametric assumptions (Hu et al., 2020, 2021) and improve the variable selection results (Bleich et al., 2014; Mazumdar et al., 2020; Ungaro et al., 2020; Hu et al., 2021, 2020; Ji et al., 2020). The variable selection methods based on nonparametric modeling techniques are primarily focused on the complete data. For variable selection on incomplete data, Hu et al. (2021) investigated combining bootstrap imputation, as suggested in Long and Johnson (2015), with six variable selection methods, and found that the tree-based variable selection methods, permutation-based Bayesian additive regression trees (BART) (Bleich et al., 2014) and recursive elimination based extreme gradient boosting (XGB) (Chen and Guestrin, 2016), random forests (RF) (Breiman, 2001) and conditional random forests (CRF) (Hothorn et al., 2006), had substantially better performance than two commonly used parametric methods, the backward stepwise selection (Wood et al., 2008) and lasso (Tibshirani, 1996); and all methods improve on the naive complete-case analysis for variable selection, when data are MAR. Hu et al. (2021) found that in the incomplete data with either a small or large percentage of missing values, BART and XGB had the overall best performance, which, when used with bootstrap
imputation via either MICE (Van Buuren, S and Groothuis-Oudshoorn, C.G.M, 2010) or missForest (Tang and Ishwaran, 2017), can recover the performance on the fully observed data; and that RF and CRF had closely similar performance across all scenarios.

The goal of this paper is to describe three alternative strategies for handling the variable selection problem in the presence of missing covariate and outcome data and compare them with the bootstrap imputation based methods. Although Hu et al. (2021) show the permutation based BART method used with bootstrap imputation (we term the approach as BI-BART) has the best overall performance, this method is computationally expensive as it requires both bootstrap and permutation on each bootstrap sample, and therefore is not easily scalable to large-scale healthcare data. To lessen the computational burden, we propose an inference based strategy that combines the BART model’s variable inclusion proportions on multiply imputed datasets using Rubin’s rule (Rubin, 2004) for variable selection. We refer to this strategy as RR-BART. We show via simulation that RR-BART achieves a significant reduction in computational times with the variable selection results comparable to BI-BART.

The three alternative strategies are: 1) missingness incorporated in attributes (MIA) (Kapelner and Bleich, 2015; Chen and Guestrin, 2016). MIA is a technique offered by some tree-based methods that treats the missingness as a value in its own right in the splitting rule. As MIA cannot handle missing outcome data, we implement this technique on cases with complete outcome data; 2) encoding the missing data of a discrete variable as “unknown”. This strategy creates an additional level to a discrete variable representing the missing values, which does not require any assumptions or need for imputation; 3) Using multiple imputation and Rubin’s rule to combine inferences about variable importance measures on multiply imputed datasets. Given that in Hu et al. (2021), all four tree-based methods had good performance and that RF and CRF had similar performance, we focus on BART, XGB and RF and compare them to the backward stepwise selection and lasso. Since there was no significant difference in the variable selection results due to imputation techniques between MICE and missForest (Hu et al., 2021), we use MICE throughout as the imputation method. For strategy 1), we consider MIA within BART and XGB, which can be implemented by specifying the appropriate missing data argument in each method’s R function. Strategy 2) is applied to all methods and strategy 3) applies to the BART model as we can obtain the uncertainty intervals of the variable inclusion proportions from their posterior samples. We compare these methods using a comprehensive and representative simulation study and further demonstrate the methods by applying them to a variable selection problem in the Study of Women’s Health Across the Nation (SWAN). Throughout, we consider a binary outcome, but all methods can be extended to continuous outcomes straightforwardly.

The remainder of the paper is organized as follows. Section 2 introduces the proposed method and overviews all comparison strategies and methods. In Section 3, we describe the data generating processes and summarize the comparative findings in our simulation study. Section 4 provides a case study where we investigate risk factors for 3-year metabolic syndrome among women in their middle years using data from the SWAN study. Section 5 concludes with a discussion.

2 Methods

2.1 Overview

We combine bootstrap imputation with five variable selection methods and term the methods as BI-BART, BI-XGB, BI-RF, BI-LASSO, BI-Stepwise. The bootstrap imputation based variable selection procedure consists in three steps: 1) generate $B$ (we use $B = 100$ in simulations (Section 3) and case study (Section 4) ) bootstrap datasets, 2) conduct imputation for each bootstrap dataset, 3) perform variable selection on each imputed (complete) dataset and select variables that appear in at least $\pi B$ datasets, where $\pi$ is the selection threshold with the optimal value depending on the specific method.

As discussed in the alternative strategy 1), we also implement MIA within the BART model and the XGB model while performing variable selection: MIA-BART and MIA-XGB. The idea of MIA is to modify the splitting rule when constructing a new tree branch by considering the missingness itself as a

© 2021 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
www.biometrical-journal.com
valid splitting criterion. The algorithm chooses one of the following three rules for all variables for splitting
X and all splitting values \( x_c \): 1) if \( X \) is observed and \( X \leq x_c \), send this observation left; otherwise, send
this observation right. If \( X \) is missing, send this observation left, 2) if \( X \) is observed and \( X \leq x_c \), send
this observation left; otherwise, send this observation right. If \( X \) is missing, send this observation right, 3) If \( X \) is missing, send this observation left; if it is observed, regardless of its value, send this observation
right. We refer to (Kapelner and Bleich, 2015) for a detailed description of the MIA procedure.

Our proposed inference-based variable selection method RR-BART uses the BART model as the infra-
structure and Rubin’s rule to combine the estimates and variances of the variable inclusion proportions on
multiply imputed datasets for variable selection. Section 2.2 describes key steps of our proposed approach.

We use multivariate amputation (Schouten et al., 2018) to generate missing data. This approach can
generate a general missing data pattern under the specified missing data mechanism with a desired miss-
ingness percentage, which may not be appropriately controlled by the commonly used stepwise univariate
amputation procedure (Schouten et al., 2018). In multivariate amputation, the complete data can be divided
into a prespecified number of subsets, to each of which a specific missing data pattern under a mechanism
can be assigned through the weighted sum score, and then the subsets can be merged together to have the
desired proportion of missingness. For example, the the weighted sum score for an MAR predictor \( X_3 \) of
a subset can follow \( wss_{x_3} = w_1 f_1 + w_2 f_2 \), where \( w_1 \) and \( w_2 \) are respectively the user-specified weights
for \( X_1 \) and \( X_2 \), on which the missingness of \( X_3 \) depends. A logistic distribution function is then ap-
plied to \( wss_{x_3} \) to compute the missingness probability for \( X_3 \). More detail of the multivariate amputation
procedure can be found in Schouten et al. (2018).

We assess the performance of variable selection using four metrics frequently used in the literature
(Wood et al., 2008; Bleich et al., 2014): i) precision, the proportion of truly useful predictors among all
selected predictors, ii) recall, the proportion of truly useful variables selected among all useful variables,
iii) \( F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \), the harmonic mean of precision and recall. The \( F_1 \) score balances between
avoiding selecting irrelevant predictors (precision) and identifying the full set of useful predictors (recall), and iv) Type I error, the mean of the probabilities that a method will incorrectly select each of the noise
predictors. We next briefly describe each method and refer to Hu et al. (2021) for a detailed review.

### 2.2 Bayesian Additive Regression Trees

BART is a tree-based Bayesian sum-of-trees model with regularizing priors to control overfitting by keep-
ing the contribution of each individual tree small (Chipman et al., 2010). BART is a likelihood-based
machine learning method, offering an additional advantage over the algorithm-based machine learning
methods of providing proper representations of uncertainty intervals via the posterior (Hu et al., 2020; Hu
and Gu, 2021; Hu et al., 2021). Details of the BART model can be found in Chipman et al. (2010). Bleich
et al. (2014) developed a permutation based variable selection method using the variable inclusion propor-
tion (VIP), the proportion of times each predictor is chosen as a splitting rule divided by the total number
of splitting rules appearing in the model. The response vector is permuted a number of times (e.g, 100) and
the BART model is fitted to each of the permuted response vectors and the original predictors; the VIPs
computed from each of BART model fits constitute the null distributions. A variable is selected if its VIP
from the BART model fitted to the unpermuted response is larger than the
\( \begin{align*}
\text{quantile of its own null & score} \end{align*} \)

We propose an inference-based approach, RR-BART. The key steps of RR-BART approach are:

1. **Impute the data** \( M \) times; fit a BART model to each imputed dataset and draw \( P \) Markov chain Monte
   Carlo (MCMC) posterior samples of the VIP for a predictor \( X_k \), \( k = 1, \ldots, K \). We then have a
distribution of \( \text{VIP}_{m,p} \), \( m = 1, \ldots, M \) \( p = 1, \ldots, P \).
2. Calculate the average VIP across \( P \) posteriors and \( M \) imputed datasets for each predictor \( \text{VIP}_{k..} = \frac{1}{M} \sum_{m=1}^{M} \min_{k=1,\ldots,K} (\text{VIP}_{km}) \); calculate the distance between each VIP score and the minimum average, \( \Delta_{kmp} = \text{VIP}_{kmp} - \text{VIP}_{k..} \) for \( k = 1,\ldots,K, m = 1,\ldots,M, p = 1,\ldots,P \).

3. Apply the Rubin’s rule (Rubin, 2004; Hu and Hogan, 2019) to the distribution of \( \Delta_{kmp} \). The overall mean and total variance of the distance score for predictor \( X_k \) are calculated as follows:

\[
\bar{Q}_k = \frac{1}{M} \sum_{m,p} \Delta_{kmp} / MP,
\]

Within-imputation variance \( W_k = \frac{1}{M} \sum_{m} \text{Var}(\bar{\Delta}_{km}.), \)

Between-imputation variance \( B_k = \frac{1}{M-1} \sum_{m} (\bar{\Delta}_{km}. - \bar{\Delta}_{k..})^2, \)

Total variance \( T_k = W_k + (1 + \frac{1}{M})B_k, \)

where \( \text{Var}(\bar{\Delta}_{km}.). \) is the variance among \( \{ \Delta_{kmp}, p = 1,\ldots,P \} \) divided by the sample size \( n \), \( \bar{\Delta}_{km}. = \frac{1}{P} \sum_{p} \Delta_{kmp} \) and \( \bar{\Delta}_{k..} = \frac{1}{MP} \sum_{m} \sum_{p} \Delta_{kmp} \). The \( 1-\alpha \) confidence interval for variable \( k \)'s average distance score is \( \bar{Q}_k \pm t_{df,1-\alpha} \sqrt{T_k} \), where the degrees of freedom for the \( t \)-distribution is \( df = (M - 1)/((B_k + B_k/M)/T_k)^2 \) (Rubin, 2004).

4. Select variables with the \( 1-\alpha \) confidence intervals that do not contain zero.

### 2.3 Random forests

RF is a tree-based ensemble method (Breiman, 2001) that aggregates a distribution of decision trees, each constructed on a bootstrap sample. In the tree-building process on each bootstrap sample, a random subset of predictors is considered for each split, and certain observations are not used for fitting the tree model and are called out-of-bag (OOB) samples. The OOB data can be used to evaluate the performance of the RF model. The RF model provides the OOB variable importance score, which is calculated by permuting the variable’s values in the OOB sample and recording the difference in prediction accuracy between the permuted and original data. The OOB variable importance score is frequently used in variable selection approaches (Díaz-Uriarte and De Andres, 2006; Hu et al., 2020; Ungaro et al., 2020). We use an approach based on recursive elimination of variables (Díaz-Uriarte and De Andres, 2006) in combination with bootstrap imputation. On each imputed dataset, we start with an RF model including the full set of predictors and record the variable importance score for each predictor, and iteratively remove a proportion (we used 10% in simulations and case study) of predictors with the smallest variable importance scores and refitting a new forest with the remaining predictors. The OOB error rates from all the fitted RF are recorded. We select the smallest number of predictors from an iteration whose OOB error rate is within one standard error of the minimum error rate of all RF models.

### 2.4 Extreme gradient boosting

At the core of the XGB method is gradient boosting (Chen and Guestrin, 2016). Boosting is a process in which a weak learner is boosted into a strong learner. The idea of gradient boosting is to build a model via additive functions that minimizes the loss function (exponential loss for classification), which measures the difference between the predicted and observed outcomes (Friedman et al., 2000). XGB uses a gradient tree boosting model with each additive function being a decision tree mapping an observation to a terminal node. In addition, XGB uses shrinkage and column subsampling to further prevent overfitting (Chen and
Shrinkage scales newly added weights by a factor after each step of tree boosting and the column subsampling selects a random subset of predictors for split in each step of tree boosting. As implemented in Hu et al. (2021), we use the recursive feature elimination procedure with XGB for variable selection on each imputed dataset. Because XGB does not supply OOB data, we evaluate model prediction accuracy on a 50% hold-out set.

2.5 Least absolute shrinkage and selection operator

The lasso is among the most popular shrinkage methods (Tibshirani, 1996). Shrinkage methods perform variable selection by regularizing the coefficient estimates. Variable selection via lasso for a binary outcome is by maximizing the log-likelihood of a logistic regression penalized for model complexity, which is concave and can be solved using nonlinear programming methods. The $\ell_1$ penalty in the penalized version of the log-likelihood has an effect of forcing some parameter coefficients to be exactly equal to zero when the tuning parameter $\lambda$ is sufficiently large. The variables with nonzero coefficients are then selected. The optimal $\lambda$ can be selected using cross validation. In our simulations, we selected $\lambda$ as the largest value such that the mean cross-validated deviance was within one standard error of the minimum deviance.

2.6 Backward Stepwise Selection

We consider the backward stepwise selection procedure based on statistical testing as implemented in Wood et al. (2008) and Hu et al. (2021). The procedure starts with a model including all potential predictors. The backward selection process deletes from the model the predictors that have the least impact on the fit (at a pre-specified statistical significance level $\alpha_1$) and the forward selection process may add back certain removed variables into the model according to a pre-specified statistical significance level ($\alpha_2$). In our simulations and case study, we used $\alpha_1 = \alpha_2 = 0.05$.

3 Simulation

The simulations are motivated by information and data structures observed in our case study in Section 4. We considered a sample size of $n = 1000$ and generated 15 predictors, among which five ($X_1, \ldots, X_5$) were truly associated with the outcome (useful predictors) and 10 were noise predictors ($X_6, \ldots, X_{15}$). Predictors $X_1$, $X_2$ and $X_3$ were generated independently from the standard normal distribution $N(0,1)$, and $X_4$ and $X_5$ were simulated from conditional distributions given $x_1, x_2$ and $x_3$ and were amputated to have missing values under the MAR mechanism. The true data models for $X_4$ and $X_5$ are:

$$x_4 \mid x_1, x_2, x_3 \sim N(0.5x_1 + 0.5x_2 + 0.3x_3^2 + 0.3x_1x_2, 1)$$

$$x_5 \mid x_1, x_2, x_3 \sim \text{Bern}(\logit^{-1}(-0.3 + 0.3x_1^2 + 0.5x_2 + 0.5x_3 + 0.3x_1x_3)).$$

The noise predictors $X_6, \ldots, X_{10}$ were independently generated from $N(0,1)$, and $X_{11}, \ldots, X_{15}$ from Bern (0.5). The outcome is related to the useful predictors as follows:

$$\Pr(y = 1 \mid x_1, \ldots, x_5) = \logit^{-1}(−4.4 + 0.1x_1^2 + 4.5\sin(0.1\pi x_2x_3) + x_4 + 1.8x_5 + 0.6x_1x_5)$$

For a fair comparison of methods, the outcome model is designed to make it difficult for any method to model the true covariate-outcome relationship. We include both linear and nonlinear forms of continuous predictors ($x_4$ and $x_1$), discrete predictors ($x_5$) and nonaddictive effects ($x_2x_3$).

Following generating the full data, we amputated $X_4$, $X_5$ and $Y$ under the MAR mechanism using the multivariate amputation procedure (Schouten et al., 2018; Hu et al., 2021). Our desired proportion of missingness in outcome $Y$ is 35% and in the entire dataset is 70%. This is motivated by the real-world SWAN data used in our case study. In SWAN data, the proportion of missing values in each covariate can be as small as 0.1% and no greater than 27%, but the proportion of complete cases is only 39%. The large
overall percentage of missingness with small-to-moderate proportion of missing values in each variable is frequently encountered in large-scale health datasets. In fact, previous simulation studies demonstrate that multiple imputation could provide unbiased results when the proportion of missing data is up to 90% (Madley-Dowd et al., 2019; Hughes et al., 2019). We first randomly split the complete data into 6 subsets with the following proportions of cases: 0.47, 0.25, 0.25, 0.01, 0.01, and 0.01. Each subset was then amputated by specifying the weighted sum score and applying the logistic distribution function to the score. Web Section 1 in the Supplementary Materials describes the multivariate amputation procedure.

There are eight methods for variable selection under investigation: BI-BART, BI-RF, BI-XGB, BI-LASSO, BI-Stepwise, MIA-BART, MIA-XGB, and RR-BART. Each method is implemented with two strategies for handling missing data in discrete variables: imputation and creating an additional category for missing data. Complete data variable selection methods are also applied to the full data (before amputation) and complete cases. When implementing imputation via MICE, all predictors available to the analyst will be included in the imputation model. For bootstrap imputation based methods, we draw $B = 100$ bootstrap samples and evaluate the performance metrics among 200 Monte Carlo replications. The performance of our proposed method RR-BART is assessed among 200 replications, each with 40 imputed datasets. The performance of strategies without imputation (MIA, full data and complete cases) is evaluated among 1000 replications. The multivariate amputation procedure was conducted using the `ampute` function in the R package mice (Schouten et al., 2018). Imputation was performed using the R package mice (Van Buuren, S and Groothuis-Oudshoorn, C.G.M, 2010). Bootstrap imputation based methods are implemented following the codes provided with Hu et al. (2021). R codes to implement our proposed method RR-BART and MIA within BART or XGB, as well as to replicate our simulation studies will be provided in the GitHub page of the corresponding author upon acceptance of this manuscript.

Table 1 displays the performance results of all variable selection methods on the full data, incomplete data and among complete cases. On the fully observed data, all three tree-based machine learning methods, BART, RF and XGB, outperform the two parametric methods, LASSO and backward stepwise selection, with RF boasting the highest $F_1$ score and XGB tending to select more noise predictors (higher Type 1 error). All methods have substantially deteriorated performance amongst complete cases. Combining with strategies for handling missing data, all methods except backward stepwise selection can recover the performance achieved on fully observed data. However, for BART and XGB, implementing MIA within the models gives lower precision, recall and $F_1$ and higher type I error than combining bootstrap imputation with the models. Our proposed method RR-BART has a similarly good overall performance (judged by $F_1$) but boasts a lower type I error compared to BI-BART with the optimal $\pi$, and when compared to MIA-BART, RR-BART has a better performance suggested by a higher recall and $F_1$ score.

Presented in Table 2 is a side-by-side comparison of six variable selection methods performed on incomplete data with two strategies for handling a missing at random discrete variable: one is imputing the missing values and one is creating a separate category of “unknown” for the missing values. The results suggest that all methods except BI-LASSO have similar or better performance when the missing values of a discrete variable are treated as a separate category without imputation. For the LASSO, setting missing values as unknown considerably increases recall but decreases precision, due to substantially increased type I error (more noise predictors selected).

Figure 1 plots the distributions of the numbers of selected noise predictors and useful predictors for each of seven methods across 200 replications. Two observations can be made. First, nonparametric machine learning methods, BI-BART, RR-BART, BI-RF and BI-XGB, tend to select more useful predictors and less noise predictors than parametric methods, BI-LASSO and BI-Stepwise. Second, across the four machine learning methods, BI-XGB selects more noise predictors, corresponding to its higher type I error and lower precision. Our proposed approach RR-BART is among the three methods that tend to select the least noise predictors. BI-RF selects the most number of useful predictors, followed by RR-BART.

A close examination of the power for each useful predictor is shown in Web Table 1. The power is calculated as the proportion of times a variable is selected across 200 replications. The average power across all useful predictors is equal to recall. When using imputation for both $X_4$ and $X_5$, BI-BART and BI-XGB
Table 1  Simulation results for each variable selection approach performed on the full data, incomplete data and complete cases. For bootstrap imputation based methods on incomplete data, we show results corresponding to the best threshold values of $\pi$ (based on $F_1$) as well as results on the complete cases.

|                      | Precision | Recall | $F_1$ | Type I error |
|----------------------|-----------|--------|-------|--------------|
| **Full data**        |           |        |       |              |
| BART                 | 1.00      | 0.78   | 0.87  | 0.00         |
| RF                   | 0.97      | 0.95   | 0.95  | 0.02         |
| XGB                  | 0.86      | 0.91   | 0.87  | 0.11         |
| LASSO                | 1.00      | 0.47   | 0.63  | 0.00         |
| Stepwise             | 0.88      | 0.56   | 0.67  | 0.05         |

| **Incomplete data**  |           |        |       |              |
| RR-BART              | 0.99      | 0.84   | 0.90  | 0.00         |
| BI-BART $\pi = 0.1$  | 0.99      | 0.80   | 0.87  | 0.01         |
| MIA-BART             | 1.00      | 0.72   | 0.83  | 0.00         |
| BI-RF $\pi = 0.3$    | 0.99      | 0.98   | 0.99  | 0.00         |
| BI-XGB $\pi = 0.5$   | 0.92      | 0.88   | 0.89  | 0.05         |
| MIA-XGB              | 0.88      | 0.84   | 0.83  | 0.09         |
| BI-LASSO $\pi = 0.5$ | 0.90      | 0.55   | 0.67  | 0.04         |
| BI-Stepwise $\pi = 0.7$ | 0.90 | 0.46   | 0.60  | 0.03         |

| **Complete cases**   |           |        |       |              |
| BART                 | 0.99      | 0.52   | 0.67  | 0.00         |
| RF                   | 0.98      | 0.56   | 0.69  | 0.01         |
| XGB                  | 0.81      | 0.66   | 0.69  | 0.11         |
| LASSO                | 0.98      | 0.24   | 0.48  | 0.00         |
| Stepwise             | 0.85      | 0.47   | 0.59  | 0.05         |

have subpar performance in identifying the discrete variable $X_5$ and good ability of detecting nonlinear ($X_1$) and nonadditive forms ($X_2$ and $X_3$) of continuous variables. On the contrary, parametric methods BI-LASSO and BI-Stepwise have a poor performance in selecting continuous predictors of complex forms and good performance in identifying discrete variables. Our proposed method RR-BART, though has a lower power in selecting nonadditive predictors ($X_2$ and $X_3$) than BI-BART, significantly improves the power for the discrete variable $X_5$, which leads to a better average power, or recall, compared to BI-BART. Interestingly, using the strategy of creating an additional category for the missing values of $X_5$ significantly improves the power of BI-BART in detecting $X_5$ and in turn the recall. The improvement is also observed for RR-BART, but not for BI-XGB and BI-LASSO.
Table 2 Comparisons of variable selection results of six methods performed on incomplete data with two strategies for handling a missing at random discrete variable: one is imputing the missing values and one is creating a separate category of “unknown” for the missing values. For bootstrap imputation based methods, results corresponding to the best threshold values of $\pi$ (based on $F_1$) are shown.

| Method      | Imputation Type | Precision | Recall | $F_1$ | Type I error | Setting missing values as unknown Type I error |
|-------------|----------------|-----------|--------|-------|--------------|-----------------------------------------------|
| RR-BART     |                | 0.99      | 0.84   | 0.90  | 0.00         | RR-BART                                      |
| BI-BART $\pi = 0.1$ |            | 0.99      | 0.80   | 0.87  | 0.01         | BI-BART $\pi = 0.1$                           |
| BI-RF $\pi = 0.3$    |            | 0.99      | 0.98   | 0.99  | 0.00         | BI-RF $\pi = 0.3$                             |
| BI-XGB $\pi = 0.5$  |            | 0.92      | 0.88   | 0.89  | 0.05         | BI-XGB $\pi = 0.5$                            |
| BI-LASSO $\pi = 0.5$ |            | 0.90      | 0.55   | 0.67  | 0.04         | BI-LASSO $\pi = 0.5$                           |
| BI-Stepwise $\pi = 0.7$ |            | 0.90      | 0.46   | 0.60  | 0.03         | BI-Stepwise $\pi = 0.7$                       |

Finally, it has been shown in the literature that the BART model for binary outcomes may be sensitive to the choice of prior for the hyperparameter $k$ (Dorie et al., 2016; Hu et al., 2020). We conducted a sensitivity analysis in Web Section 2 of the Supplementary Materials to assess whether the specification

Figure 1 Distributions of the numbers of selected noise predictors and useful predictors for each of six methods across 200 replications. The total number of useful predictors is 5 and the total number of noise predictors is 10. MIA-BART and MIA-XGB are not shown for their inferior performance compared to imputation based methods.
for $k$ impacts the performance of BART based methods, BI-BART and RR-BART. Web Figure 1 suggests that both methods may be moderately sensitive to $k$ with respect to precision and type I error, but relatively insensitive to $k$ in terms of recall and $F_1$. Web Figure 2 demonstrates that both methods are insensitive to $k$ when selecting continuous variables in either the linear form ($X_4$) or nonlinear form ($X_1$), however, our proposed approach RR-BART is less sensitive to the amount of shrinkage via $k$ than BI-BART in selecting the discrete variable $X_5$ and nonadditive continuous variables $X_2$ and $X_3$.

4 Case study

We demonstrate the methods by addressing a variable selection problem using data from the SWAN study. The SWAN study was a multicenter, longitudinal study in the US with primary aim of understanding women’s health across the menopause transition. The SWAN study enrolled 3305 women aged between 42 and 52 in 1996-1997 and followed them up to 2018 annually. A detailed description of the study can be found in Janssen et al. (2008). The goal of this case study is to identify predictors of 3-year incidence of metabolic syndrome, which is an important health research problem but has been less studied in the literature for women during their middle years.

Metabolic syndrome is a cluster of conditions that occur together and has been shown to increase the risk of heart disease, stroke and type 2 diabetes (Kazlauskaite et al., 2020). Following prior work (Kazlauskaite et al., 2020), we define metabolic syndrome as the presence of at least three of the following five symptoms: abdominal obesity, hypertension, hypertriglyceridemia, impaired fasting glucose, and low high-density lipoprotein cholesterol level. Our analysis included 2313 women who did not have metabolic syndrome at enrollment. Among these women, 251 (10.9%) developed metabolic syndrome within three years of enrollment, 1240 (53.6%) did not, and the remaining 822 (35.5%) did not respond to the survey questions giving rise to missing outcomes. We considered 15 potential predictors (3 binary and 12 continuous variables) including socioeconomic status, life style and biomarkers based on the previous literature (Janssen et al., 2008; Kazlauskaite et al., 2020). The name and definition of the 15 variables are displayed in Table 3. Nine out of the 15 potential predictors have missing data with the proportion of missingness ranging from 0.1% to 27.1%. Only 39.2% observations have observed data in both covariates and outcomes, i.e., the overall percentage of missingness in the data is 60.8%.

Given that the four nonparametric methods BI-RF, RR-BART, BI-BART and BI-XGB have substantially better performance than other approaches considered, we applied these four methods to the SWAN data to select important predictors for 3-year incidence of metabolic syndrome. For BI-RF, BI-BART and BI-XGB, we respectively used the optimal threshold value $\pi = 0.3$, $\pi = 0.1$ and $\pi = 0.5$, suggested for each method by the simulation study. We also created an additional category for the missing values of the discrete variables for BI-BART, RR-BART and BI-RF given that our simulation study demonstrates that their performance with this strategy is better than their performance using imputation for discrete variables.

All four methods selected the same six predictors: diastolic blood pressure (DIABP), high density lipoprotein cholesterol (HDLRESU), systolic blood pressure (SYSBP), tissue plasminogen activator (TPA), triglycerides (TRIGRES) and waist circumference (WAIST). The results are in accordance with prior work on risk factors for metabolic syndrome (Carnethon et al., 2004; Wei et al., 2018). For example, blood pressure, cholesterol level, waist circumstance and triglycerides have been identified as common risk factors for metabolic syndrome. Our analysis also selected TPA as a potentially important risk factor. This may be justified in the clinical context: TPA is associated with insulin resistance, which was found to be involved in the pathogenesis of impaired fasting gluoces (Rao et al., 2004).

5 Discussion

We investigate strategies for variable selection with missing covariate and outcome data in large-scale health study. Prior work has shown that combining bootstrap imputation with tree-based variable selection
Table 3 Names and definitions of variables used in the case study in Section 4.

| Variable   | Definition                                                                 |
|------------|---------------------------------------------------------------------------|
| DEGREE     | Education degree                                                          |
| DIABP      | Diastolic blood pressure                                                  |
| DTTRANS    | Trans fats, grams (g)                                                     |
| HDLRESU    | High density lipoprotein cholesterol, milligrams per deciliter (mg/dl)   |
| HPBMD      | Total hip bone mineral density                                            |
| LDLRESU    | Low density lipoprotein cholesterol, milligrams per deciliter (mg/dl)    |
| LPA1RES    | Lipoprotein Lp(a), milligrams per deciliter (mg/dl)                       |
| PAIRESU    | PAI-1, nanograms per milliliter (ng/ml)                                  |
| PHYSWORK   | Work physical compared to other women same age                            |
| SMOKERE    | Ever smoked regularly                                                     |
| SPBMD      | Total spine bone mineral density                                          |
| SYSBP      | Systolic blood pressure                                                   |
| TPA        | Tissue plasminogen activator, nanograms per milliliter (ng/ml)            |
| TRIGRES    | Triglycerides, milligrams per deciliter (mg/dl)                           |
| WAIST      | Waist circumference (cm)                                                  |

methods has good operating characteristics with respect to selecting the most relevant predictors (Hu et al., 2021). The computational cost of this method, however, can be high in large-scale data, given that both bootstrap and recursive nonparametric modeling are needed. Here we propose an inference-based method that leverages the underlying Bayesian probability model of the tree-based machine learning technique BART, and uses Rubin’s rule to combine estimates and variances of the VIPs of each predictor provided by the BART model on multiply imputed datasets for variable selection. In addition, we consider two alternative strategies other than imputation for handling missing data: treating the missing values of discrete variables as an additional level and implementing MIA within BART and XGB, which accommodate missing covariates by modifying the binary splitting rules.

Our simulation study suggests that the proposed method RR-BART performs at least as well as bootstrap imputation based BI-BART and can recover the performance on the fully observed data, even with a large percentage of overall missingness and in complex conditions of nonlinearity and nonadditivity. RR-BART is less sensitive to the end note prior via the hyperparameter $k$ than BI-BART in detecting discrete variables or nonadditive continuous variables. Creating a separate level for the missing values in a binary variable significantly improves the power of BI-BART in selecting the discrete variable compared to imputing the missing values. This strategy is also beneficial for RR-BART and BI-RF. This may be because the additional category representing the missing data itself adds information to be considered for splitting in the tree-constructing process. The benefit of using this strategy particularly with BI-BART or RR-BART may be more pronounced when a dataset has many binary variables. The other imputation-free strategy, MIA within BART or XGB, however has inferior variable selection performance, suggested by decreased precision and recall and increased type I error. The MIA strategy cannot accommodate missing outcome data, and therefore may be less efficient than imputation based methods due to reduced sample size.
The BI-RF method has the best performance. Hu et al. (2021) shows in their simulations that although BI-RF has a good performance, it tends to have relatively lower recall and $F_1$ score than BI-BART. However, in our simulation settings, BI-RF has a higher recall and $F_1$ than BI-RR. This is likely due to the smaller numbers of useful and noise predictors considered in our study coupled with the sensitivity of BI-BART to $k$. The performance of our proposed approach, RR-BART, is on par with BI-RF, but the computational savings offered by RR-BART are substantial. All simulations were run in R on an iMAC with a 4 GHz Intel Core i7 processor. On a dataset of size $n = 1000$ with 15 predictors, each RR-BART or BI-XGB implementation took 2 minutes to run, while each BI-RF implementation took about 8 minutes and each BI-BART took about 3 hours to run. More importantly, unlike the bootstrap imputation based methods, RR-BART does not require a selection threshold $\pi$, whose optimal value varies by methods and simulation settings.

Although our proposed method RR-BART provides promising performance for variable selection in the presence of MAR covariate and outcome data, it is possible to further improve the method for the big-$n$-small-$p$ situation, which is frequently encountered in health registry data. Another possible research avenue is to derive an alternative nonparametric measure of variable importance rather than the VIP that reflects the impact of inclusion or deletion of predictors on the model prediction accuracy in the presence of missing data (Williamson et al., 2021). Finally, extending the methods to accommodate MNAR data could be a worthwhile contribution.

**Acknowledgements** This work was supported in part by award ME_2017C3_9041 from the Patient-Centered Outcomes Research Institute, and by grant R21CA245855 from the National Cancer Institute.

**Conflict of Interest**

*The authors have declared no conflict of interest.*
References

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

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Carnethon, M. R., Loria, C. M., Hill, J. O., Sidney, S., Savage, P. J., and Liu, K. (2004). Risk factors for the metabolic syndrome: the Coronary Artery Risk Development in Young Adults (CARDIA) study, 1985–2001. *Diabetes Care* **27**, 2707–2715.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

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

Diaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine* **35**, 3453–3470.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics* **28**, 337–407.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection—a review and recommendations for the practicing statistician. *Biometrical Journal* **60**, 431–449.

Hocking, R. R. (1976). A biometrics invited paper: the analysis and selection of variables in linear regression. *Biometrics* pages 1–49.

Hogan, J. W., Daniels, M. J., and Hu, L. (2014). A bayesian perspective on assessing sensitivity to assumptions about unobserved data. In Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G., editors, *Handbook of Missing Data Methodology*, chapter 18, pages 405–434. CRC Press, Boca Raton, FL.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* **15**, 651–674.

Hu, L. and Gu, C. (2021). Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes. *Health Services and Outcomes Research Methodology* In press.

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical Methods in Medical Research* **29**, 3218–3214.

Hu, L. and Hogan, J. W. (2019). Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics* **75**, 695–707.

Hu, L., Hogan, J. W., Mwangi, A. W., and Siika, A. (2018). Modeling the causal effect of treatment initiation time on survival: Application to HIV/TB co-infection. *Biometrics* **74**, 703–713.

Hu, L., Ji, J., and Li, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Statistics in Medicine* In press.

Hu, L., Ji, J., Li, Y., Liu, B., and Zhang, Y. (2021). Quantile regression forests to identify determinants of neighborhood stroke prevalence in 500 cities in the USA: implications for neighborhoods with high prevalence. *Journal of Urban Health* **98**, 259–270.

Hu, L., Li, L., and Ji, J. (2020). Machine learning to identify and understand key factors for provider-patient discussions about smoking. *Preventive Medicine Reports* **20**, 101238.

Hu, L., Lin, J.-Y. J., and Ji, J. (2021). Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *arXiv preprint arXiv:2104.02769*.

Hu, L., Lin, J.-Y. J., Sigel, K., and Kale, M. (2021). Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Annals of Epidemiology* In press.

Hu, L., Liu, B., Ji, J., and Li, Y. (2020). Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *Journal of the American Heart Association* **9**, e016745.

Hu, L., Liu, B., and Li, Y. (2020). Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: a bayesian machine learning approach. *Preventive Medicine* **141**, 106240.
Hughes, R. A., Heron, J., Sterne, J. A., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology* **48**, 1294–1304.

Janssen, I., Powell, L. H., Crawford, S., Lasley, B., and Sutton-Tyrrell, K. (2008). Menopause and the metabolic syndrome: the Study of Women’s Health Across the Nation. *Archives of Internal Medicine* **168**, 1568–1575.

Ji, J., Hu, L., Liu, B., and Li, Y. (2020). Identifying and assessing the impact of key neighborhood-level determinants on geographic variation in stroke: a machine learning and multilevel modeling approach. *BMC Public Health* **20**, 1–12.

Kapelner, A. and Bleich, J. (2015). Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics* **43**, 224–239.

Kazlauskaite, R., Janssen, I., Wilson, R. S., Appelhans, B. M., Evans, D. A., Arvanitakis, Z., El Khoudary, S. R., and Kravitz, H. M. (2020). Is midlife metabolic syndrome associated with cognitive function change? The Study of Women’s Health Across the Nation. *The Journal of Clinical Endocrinology & Metabolism* **105**, e1093–e1105.

Little, R. J., D’Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* **367**, 1355–1360.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons, New York, 3rd edition.

Long, Q. and Johnson, B. A. (2015). Variable selection in the presence of missing data: resampling and imputation. *Biostatistics* **16**, 596–610.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology* **110**, 63–73.

Mazumdar, M., Lin, J.-Y., Zhang, W., Li, L., Liu, M., Dharmarajan, K., Sanderson, M., Isola, L., and Hu, L. (2020). Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by oncology care model (OCM) data. *BMC Health Services Research* **20**, 350.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

NATIONAL RESEARCH COUNCIL (2010). The prevention and treatment of missing data in clinical trials. The National Academies Press, Washington, DC.

Rao, S. S., Disraeli, P., and McGregor, T. (2004). Impaired glucose tolerance and impaired fasting glucose. *American Family Physician* **69**, 1961–1968.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York.

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation* **88**, 2909–2930.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* **338**.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **10**, 363–377.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288.

Ungaro, R. C., Hu, L., Ji, J., Nayar, S., Kugathasan, S., Denson, L. A., Hyams, J., Dubinsky, M. C., Sands, B. E., and Cho, J. H. (2020). Machine learning identifies novel blood protein predictors of penetrating and strictureing complications in newly diagnosed paediatric crohn’s disease. *Alimentary Pharmacology & Therapeutics* **53**, 281–290.

Van Buuren, S and Groothuis-Oudshoorn, C.G.M (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**, 1–67.

Wei, B., Liu, Y., Lin, X., Fang, Y., Cui, J., and Wan, J. (2018). Dietary fiber intake and risk of metabolic syndrome: A meta-analysis of observational studies. *Clinical Nutrition* **37**, 1935–1942.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **77**, 9–22.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine* **27**, 3227–3246.
Web-based Supplementary Materials for “Strategies for variable selection in large-scale healthcare database studies with missing covariate and outcome data”

Jung-Yi Lin¹, Liangyuan Hu∗², Chuyue Huang³, Steven Lawrence¹, and Usha Govindarajulu¹

¹Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, New York 10029
²Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, New Jersey 08854
³Primary Research Solution LLC. 115 W 18th St, New York, New York 10011

S1 Multivariate amputation in Section 3

After generating the complete data, we employed the multivariate amputation procedure to amputate $X_4$, $X_5$ and $Y$ under the MAR mechanism and the final incomplete dataset has 35% of missingness in the outcomes and 70% of missingness in the entire dataset. We first randomly split the complete data into 6 subsets with the following proportions of cases: 0.47, 0.25, 0.25, 0.01, 0.01, and 0.01. Each subset was then amputated with a specific missing data pattern as follows:

1. $wss_{y,i} = 8x_1 + 8x_2 + 8x_3 + 10x_4 + 12x_5 + 8x_2x_3 + 8x_1x_4$
2. $wss_{x_4,i} = 8x_1 + 8x_2 + 6x_3^2 + 8x_1x_2$
3. $wss_{x_5,i} = 6x_1^2 + 4x_1 + 8x_2 + 8x_3 + 8x_1x_3$
4. $wss_{y,x_4,i} = 10x_1 + 8x_2 + 8x_3$
5. $wss_{y,x_5,i} = 8x_1^2 + 6x_1 + 10x_2 + 8x_3 + 8x_1x_3$
6. $wss_{y,x_4,x_5,i} = 8x_1 + 8x_2 + 8x_3^2 + 6x_3 + 8x_1x_2$.

In subsets (4)-(6), the joint missingness in $Y$ and $X_4$, $Y$ and $X_5$ and $Y$, $X_4$ and $X_5$ was created. A right-tailed type of missingness was applied in subsamples (1)-(3) and a centered type of missingness was applied in subsamples (4)-(6).

S2 Sensitivity of BI-BART and RR-BART to $k$

We conducted a sensitivity analysis to assess whether the choice of end-node ($\mu$) prior, specifically via the hyperparameter $k$, impacts the performance of BART based methods, BI-BART and RR-BART. We consider four values $k = 1, 2, 3, 5$ reflecting moderate to heavy shrinkage for the $\mu$ prior hyperparameter, and evaluate the performance of BI-BART and RR-BART using the four performance metrics: precision, recall, $F_1$ and type I error, as well as the power for each of five useful predictors. Suggested by Web Figure 1, both methods are moderately sensitive to $k$ with

*Corresponding author: e-mail: liangyuan.hu@rutgers.edu, Phone: 732-235-4646, Fax: 732-235-5476
respect to precision and type I error, but relatively insensitive to \( k \) in terms of recall and \( F_1 \). RR-BART achieves the best \( F_1 \) with \( k = 2 \) while a slightly heavier shrinkage via \( k = 3 \) seems to marginally improve the \( F_1 \) score for BI-BART. Web Figure 2 takes a closer look at how sensitive the power for each of five useful predictors is to the specification of \( k \). Both methods are insensitive to \( k \) in identifying a continuous predictor when it is in either the linear (\( X_4 \)) or nonlinear (\( X_1 \)) form. With discrete variables or continuous variables in conditions of nonadditivity, both methods are sensitive to \( k \), but the level of shrinkage has a smaller impact on RR-BART than on BI-BART. For BI-BART, the identification of the discrete variable \( X_5 \) requires a heavy shrinkage but the identification of the nonadditive predictors (\( X_2 \) and \( X_3 \)) tend to favor a moderate shrinkage. For RR-BART, moderate shrinkage via \( k = 2 \) or \( k = 3 \) works well for both discrete variables and nonadditive continuous variables.

S3 Additional tables and figures

Web Table 1: Power for each of five useful predictors of six methods performed on incomplete data with two strategies for handling a missing at random discrete variable: one is imputing the missing values and one is creating a separate category of “unknown” for the missing values. Power is calculated as the proportion of times a variable is selected across 200 replications. For bootstrap imputation based methods, results corresponding to the best threshold values of \( \pi \) (based on \( F_1 \)) are shown. MIA-BART and MIA-XGB are not shown for their inferior performance compared to imputation based methods.

|                  | \( X_1 \) | \( X_2 \) | \( X_3 \) | \( X_4 \) | \( X_5 \) |
|------------------|-----------|-----------|-----------|-----------|-----------|
| \begin{tabular}{l} \textbf{Imputation} \\ \text{BI-BART}  \\ \text{\( \pi = 0.1 \)} \end{tabular} | 1.00 | 0.85 | 0.89 | 1.00 | 0.25 |
| \begin{tabular}{l} \textbf{Imputation} \\ \text{RR-BART} \end{tabular} | 1.00 | 0.62 | 0.67 | 1.00 | 0.93 |
| \begin{tabular}{l} \textbf{Imputation} \\ \text{BI-RF}  \\ \text{\( \pi = 0.3 \)} \end{tabular} | 1.00 | 0.99 | 1.00 | 1.00 | 0.93 |
| \begin{tabular}{l} \textbf{Imputation} \\ \text{BI-XGB}  \\ \text{\( \pi = 0.5 \)} \end{tabular} | 1.00 | 0.91 | 0.92 | 1.00 | 0.60 |
| \begin{tabular}{l} \textbf{Imputation} \\ \text{BI-LASSO}  \\ \text{\( \pi = 0.5 \)} \end{tabular} | 0.35 | 0.28 | 0.14 | 1.00 | 1.00 |
| \begin{tabular}{l} \textbf{Imputation} \\ \text{BI-Stepwise}  \\ \text{\( \pi = 0.7 \)} \end{tabular} | 0.12 | 0.08 | 0.08 | 1.00 | 1.00 |

|                  | \( X_1 \) | \( X_2 \) | \( X_3 \) | \( X_4 \) | \( X_5 \) |
|------------------|-----------|-----------|-----------|-----------|-----------|
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{BI-BART}  \\ \text{\( \pi = 0.1 \)} \end{tabular} | 1.00 | 0.86 | 0.89 | 1.00 | 0.97 |
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{RR-BART} \end{tabular} | 1.00 | 0.72 | 0.76 | 1.00 | 0.96 |
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{BI-RF}  \\ \text{\( \pi = 0.3 \)} \end{tabular} | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 |
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{BI-XGB}  \\ \text{\( \pi = 0.5 \)} \end{tabular} | 1.00 | 0.94 | 0.95 | 1.00 | 0.31 |
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{BI-LASSO}  \\ \text{\( \pi = 0.5 \)} \end{tabular} | 0.39 | 0.56 | 0.37 | 1.00 | 0.77 |
| \begin{tabular}{l} \textbf{Setting missing values as unknown} \\ \text{BI-Stepwise}  \\ \text{\( \pi = 0.7 \)} \end{tabular} | 0.11 | 0.13 | 0.08 | 1.00 | 1.00 |
Web Figure 1: The sensitivity of BI-BART and RR-BART’s precision, recall, $F_1$ and type I error to $k$. Four values $k = 1, 2, 3, 5$ are considered reflecting moderate to heavy shrinkage for the $\mu$ prior hyperparameter.
Web Figure 2: The sensitivity of the power of BI-BART and RR-BART to $k$ for each of five useful predictors. Four values $k = 1, 2, 3, 5$ are considered reflecting moderate to heavy shrinkage for the $\mu$ prior hyperparameter.