Inference from Non-Random Samples Using Bayesian Machine Learning*

Yutao Liu, Andrew Gelman, Qixuan Chen†

10 Oct 2021

Abstract

We consider inference from non-random samples in data-rich settings where high-dimensional auxiliary information is available both in the sample and the target population, with survey inference being a special case. We propose a regularized prediction approach that predicts the outcomes in the population using a large number of auxiliary variables such that the ignorability assumption is reasonable and the Bayesian framework is straightforward for quantification of uncertainty. Besides the auxiliary variables, inspired by Little & An (2004), we also extend the approach by estimating the propensity score for a unit to be included in the sample and also including it as a predictor in the machine learning models. We find in simulation studies that the regularized predictions using soft Bayesian additive regression trees yield valid inference for the population means and coverage rates close to the nominal levels. We demonstrate the application of the proposed methods using two different real data applications, one in a survey and one in an epidemiologic study.

*To appear in *Journal of Survey Statistics and Methodology*.

†Yutao Liu is a Senior Biostatistician II at Vertex Pharmaceuticals and was a PhD student in the Department of Biostatistics at Columbia University; Andrew Gelman is a Professor of Statistics and Political Science, and Qixuan Chen is an Associate Professor of Biostatistics, at Columbia University. Address correspondence to: Qixuan Chen, Department of Biostatistics, Columbia University, 722 West 168th Street, New York, NY 10032; E-mail: qc2138@cumc.columbia.edu.

Statement of significance: We propose Bayesian machine learning model-based methods for estimating population quantities using non-random samples in data-rich settings where high-dimensional auxiliary information is available both in the sample and the population. We use Bayesian Additive regression trees (BART) and soft Bayesian additive regression trees (SBART) to predict survey outcomes in the population, with the Bayesian framework straightforward for quantification of uncertainty. We further extend the trees-based estimators by also including the estimated propensity score of inclusion as a covariate to make them doubly robust to either model misspecification. Using simulation studies, we show that our proposed methods effectively reduce selection bias in the non-random sample and yield efficient estimates of population quantities, with close to the nominal level coverage rate. SBART performs better than BART and is, therefore, recommended. Including the estimated propensity score further reduces bias and improves efficiency in the BART estimator but the improvement over SBART is not obvious. However, the propensity score does provide protection against model misspecification in SBART when important predictors are omitted in the outcome models. Using our proposed methods, we can quantify and correct differences between population and a survey, a randomized trial, or an epidemiologic study.

Keywords: Bayesian machine learning; High-dimensional auxiliary variables; Non-random samples; Probability and non-probability surveys; Propensity score; Soft Bayesian additive regression trees.

1 Introduction

Inference about a target population based on sample data relies on the assumption that the sample is representative or can be adjusted to account for non-representativeness.

Unfortunately, random samples are often not available in real data problems. Therefore, there is a need to generalize inference from an available non-random sample to the target population of interest. For example, randomized controlled trials (RCTs) are considered a gold standard to estimate treatment effects, but in the presence of varying treatment effects, the measured effects can only be formally generalized to the participants within the trial. Recent evidence has indicated that subjects in RCTs can be much different from patients in routine practice. Such concern among clinicians about the external validity of RCTs has led to the underuse of effective treatments (Rothwell 2005). This highlights the importance of generalizing treatment effect of RCTs to a definable patient population.

Survey sampling is a field that specifically deals with inference on populations from samples, which can be viewed as a special case of generalizing inference. In the past several decades, large scale probability surveys have suffered increasingly high nonresponse rates. Probability surveys with low response rates are often non-representative, which challenges the validity of survey inference. In the meanwhile, recent development of information technology makes it increasingly convenient and cost-effective to collect large numbers samples with detailed information via online surveys and opt-in panels. These non-probability samples can be highly non-representative due to selection bias. Statistical methods that can effectively correct for the selection bias are needed, and can take advantage of the large amounts of auxiliary information that are typically available with such data (Baker et al. 2013).

Classical weighting methods in the survey literature such as weighting, poststratification, and raking can correct non-representativeness of survey samples with respect to a small number of discrete auxiliary variables (Deming & Stephan 1940, Valliant 1993). However, when the number of adjustment variables is large or sample size with in cells is small, such weighting methods can yield highly extreme weights and highly variable estimates of

population quantities. Alternatively, prediction using an outcome model can be used (Smith 1983). Wang, Rothschild, Goel & Gelman (2015) demonstrates, in an example of public opinion modeling during an election campaign using non-representative voter intention polls on the Xbox gaming platform, that multilevel regression and poststratification (MRP) can be used to generate accurate survey estimates from non-representative samples using population distributions of auxiliary information. Their estimates are in line with the forecasts from leading poll analysts. Kim, Park, Chen & Wu (2021) propose a mass imputation approach using generalized linear models that combines a non-random sample with a reference probability survey to obtain valid inference of population quantities. Elliott & Valliant (2017) review the pseudo-weighting and model-based prediction approaches in non-probability surveys and discussed the pros and cons of each approach.

In recent years, population data of high volume, variety, and velocity have become increasing available, with examples including administrative data or electronic medical records. Such data contain detailed individual level information and can be used to generalize inference of non-random samples to their target populations. Although poststratification, raking, and MRP methods can improve representativeness in the presence of a small number of discrete auxiliary variables, they can run into problems in high-dimensional settings. In this paper, we study the problem of inference for non-random samples when there exist external population data containing individual level auxiliary information with high dimensionality. These samples can be convenience samples or probability samples that are effectively non-random due to severe nonresponse or sampling frame undercoverage.

With high-dimensional auxiliary variables, Bayesian machine learning techniques have been shown to be effective in improving statistical inference in missing data and causal inference

(Hill 2011, Tan, Flannagan & Elliott 2019, Hahn, Murray & Carvalho 2020). Specially, Hill (2011) demonstrated the use of Bayesian additive regression trees (BART) to produce more accurate estimates of average treatment effects compared to propensity score matching, propensity-weighted estimators, and regression adjustment when the response surface is nonlinear and not parallel between treatment and control groups. Tan, Flannagan & Elliott (2019) demonstrated, in the presence of missing data, that BART reduces bias and root mean square error of doubly robust estimators when both propensity and mean models were misspecified. In survey sampling, Rafei, Flannagan & Elliott (2020) used BART to estimate pseudo-weights for non-probability samples and showed that BART provided further reduction in bias compared with conventional propensity adjustment method. Rafei, Flannagan, West & Elliott (2021) further applied this approach to extend doubly robust adjustment in non-probability surveys to protect against model misspecification. Inspired by these works, we propose Bayesian machine learning model-based methods and extensions for estimating population means using non-random samples.

Inference from non-random samples is of great importance. The methods developed in this paper can be applied not only in the context of survey inference but also in more general settings, such as RCTs and epidemiological observational studies. Using auxiliary information and the proposed Bayesian machine learning methods, we can quantify and correct differences between a population and participants in a survey, a randomized trial, or an epidemiologic study (Stuart, Cole, Bradshaw & Leaf 2011, Keiding & Louis 2016). We evaluate the proposed methods using simulation studies and demonstrate their applications in a mental health survey of Ohio Army National Guard service members and a non-random sample from an epidemiologic study using electronic medical records of COVID-19 patients.

2 Methods

2.1 Notation and background

Let U be a finite population of size N and s be a non-random sample of size n from the population. In the sample s, information on the outcome of interest y, discrete auxiliary variables $\boldsymbol{Z}$ and continuous auxiliary variables $\boldsymbol{X}$ were collected. In addition, data from the population U (e.g. census, administrative data, or electronic medical records) is also available with the same set of auxiliary variables $\boldsymbol{Z}$ and $\boldsymbol{X}$ measured for all units in the population. Figure 1 illustrates the scenario under consideration, with population data on the left and the sample data on the right. Without loss of generality, we consider a continuous variable of interest y with the estimand of interest being the finite population mean $Q(y) = \frac{1}{N}\sum_{i\in U} y_i$.

When the dimensions of $\boldsymbol{Z}$ and $\boldsymbol{X}$ are small, poststratification, raking, and MRP can be applied by first discretizing the continuous auxiliary variables $\boldsymbol{X}$ as $\boldsymbol{X}^*$ using quantiles. Using the joint distribution of discrete auxiliary variables $(\boldsymbol{Z}, \boldsymbol{X}^*)$, *poststratification* partitions the population into J disjoint poststrata with $U = \bigcup_{j=1}^{J} U_j$ of size N_j and the sample into subsamples with $s = \bigcup_{j=1}^{J} s_j$ of size n_j for the jth poststratum, correspondingly. With respect to the poststrata, the finite population mean can be rewritten as

$$Q(y) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} y_i = \frac{1}{N}\sum_{j=1}^{J} N_j\theta_j, \tag{1}$$

where $\theta_j = \frac{1}{N_j}\sum_{i\in U_j} y_i$ is subpopulation mean of poststratum j. With the assumption that the sample units in each poststratum are representative of population units in that poststratum, the poststrata means are estimated using corresponding subsample means

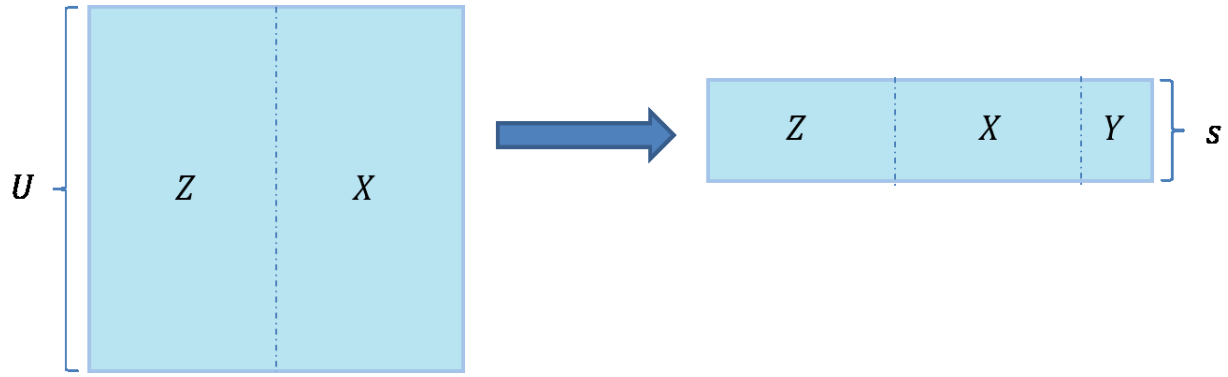

Figure 1: Population U and non-random sample s with shared discrete auxiliary variables $\boldsymbol{Z}$ and continuous auxiliary variables $\boldsymbol{X}$ as well as outcome y measured only in s.

$\hat{\theta}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i \in s_j} y_i$. The poststratification (PS) estimator takes the form

$$\widehat{Q}_{\text{PS}} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{y}_j = \frac{1}{N} \sum_{i \in s} w_i y_i, \tag{2}$$

where $w_i = N_j / n_j$ for $i \in s_j$ is the poststratification weight assigned to sample unit i in poststratum j which is inverse proportional to the sampling fraction n_j / N_j. The poststratification estimator could be numerically unstable when such partition results in small cells in the sample: small n_j and large weights w_j.

Alternatively, *raking* generates weights w_i to match successively the marginal (rather than the joint) distributions of $(\boldsymbol{Z}, \boldsymbol{X}^*)$ via iterative proportional fitting. The raking weighted estimator takes the form

$$\widehat{Q}_{\text{R}} = \frac{1}{N} \sum_{i \in s} w_i y_i, \tag{3}$$

with w_i denoting raking weights. Raking weights could be highly variable, so the resulting weighted estimators could be inefficient. Also, raking may have convergence issues as the number of auxiliary variables increases.

Gelman (2007) reviews the PS estimator from a model-based perspective, in which a

regression model is specified for the conditional distribution of the outcome given the discrete auxiliary variables: $p(y|\boldsymbol{z}, \boldsymbol{x}^*)$. For units in the same poststratum j, the auxiliary variables have the same values $(\boldsymbol{z}_j, \boldsymbol{x}_j^*)$ and, under a specified model, the outcome y has poststratum specific means $\theta_j = \mathrm{E}(y|\boldsymbol{z}_j, \boldsymbol{x}_j^*)$, $j = 1, \ldots, J$. And estimating θ_j based on the fitted model leads to the *regression and poststratification* (RP) estimator

$$\widehat{Q}_{\mathrm{RP}} = \frac{1}{N} \sum_{j=1}^{J} N_j \hat{\theta}_j, \tag{4}$$

where, $\hat{\theta}_j = \hat{\mathrm{E}}(y|\boldsymbol{z}_j, \boldsymbol{x}_j^*)$. As a special case, specifying a saturated regression model (including all possible interactions terms) allows J poststratum specific means, leading to the least-squares estimators $\hat{\theta}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i \in s_j} y_i$. As a result, $\widehat{Q}_{\mathrm{RP}} = \widehat{Q}_{\mathrm{PS}}$.

From the model-based perspective, the problem of unstable estimates due to small cells in poststratification can be viewed as a model fitting problem due to model complexity. Such perspective motivates using alternative modeling techniques to improve estimation. Instead of using classical saturated regression models, *multilevel regression and poststratification* (MRP) uses hierarchical regression models to achieve stable estimates. Both main effects and interaction terms could be specified as multilevel random effects so that information across poststrata can be partially pooled in the model fitting procedure (Gelman & Little 1997). MRP improves efficiency in the population mean estimation than poststratification and raking when data are sparse in some poststrata.

Still, it is challenging to perform MRP in high-dimensional setting, especially in the presence of a large number of noise variables not associated with y, because a parametric form needs to be specified for the multilevel regression. Also, continuous auxiliary variables need to be discretized before modeling.

The model-based RP approach can also be viewed as a prediction approach, and the RP estimator in (4) can be rewritten as

$$\widehat{Q}_{\text{RP}} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\theta}_j = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} \hat{\theta}_j = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} \widehat{\mathrm{E}}(y_{i[j]}|\boldsymbol{z}_j, \boldsymbol{x}_j^*) = \frac{1}{N}\sum_{i\in U} \widehat{\mathrm{E}}(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^*), \quad (5)$$

where $\widehat{\mathrm{E}}(y_{i[j]}|\boldsymbol{z}_j, \boldsymbol{x}_j^*)$ is a predictive value of y for unit i in poststratum j based on model $p(y|\boldsymbol{z}, \boldsymbol{x}^*)$ and the value is constant over all i within the same poststratum. This perspective motivates the use of modern statistical techniques for generalization of inference via valid predictions of the outcomes in the population. Specifically, the classical regression models in *regression and poststratification* can be replaced by any regularized prediction method that achieve stable estimates while including high-dimensional covariates. Such models also allow modeling the continuous $\boldsymbol{X}$ directly.

2.2 New Approach: Regularized Prediction

The regularized prediction approach for the data setting in Figure 1 relies on the following assumptions: (A1) unit-level auxiliary information is available for the target population of interest; (A2) the sampling mechanism is ignorable (Gelman et al. 2014), i.e., the sample inclusion indicator I and the outcome variable y are independent given the set of auxiliary variables $(\boldsymbol{Z}, \boldsymbol{X})$; and (A3) the sample and the population have the same ranges of values for the auxiliary variables. Assumption A3 cannot be satisfied by scenarios where the ranges of auxiliary variables in the population are out of ranges of those in the sample, where extrapolation is required. In addition, the performance of the regularized prediction approach largely depends on how well the prediction model is fit, and thus flexible models without clear model misspecification are desired.

2.2.1 BART and Soft BART Prediction

In the current setting, the conditional distribution of a continuous outcome given the high-dimensional auxiliary variables $p(y|\boldsymbol{z}, \boldsymbol{x})$ can be modeled using Bayesian additive regression trees (BART) (Chipman, George & McCulloch 2010) or soft Bayesian additive regression trees (SBART) (Linero & Yang 2018).

For continuous outcomes, BART and SBART assume Gaussian noise and model the location parameter using a non-parametric sum-of-trees structure, allowing both discrete and continuous auxiliary variables

$$y = G(\boldsymbol{z}, \boldsymbol{x}) + \epsilon = \sum_{m=1}^{M} g(\boldsymbol{z}, \boldsymbol{x}; T_m, \boldsymbol{\mu}_m) + \epsilon, \quad \epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \tag{6}$$

where M is fixed number of trees in the sum-of-trees structure, T_m is the m-th binary tree with $\boldsymbol{\mu}_m$ being the parameters associated with the terminal nodes, and $g(\cdot)$ is the function assigning $\boldsymbol{\mu}_m$ according to $(\boldsymbol{z}, \boldsymbol{x})$. The sum-of-trees structure naturally handles high-dimensional auxiliary variables without specifying a parametric form, accounting for categorical variables, continuous variables and possible interactions. In the Bayesian framework, quantification of uncertainty is naturally characterized by the posterior and posterior predictive distributions.

In BART, $g(\cdot)$ is a deterministic function and the potential effect of continuous predictors, either linear or nonlinear, is approximated by step functions generated by cutting the continuous predictor at various splitting points in different trees. Regularization priors are specified on $p(T_m)$, $p(\boldsymbol{\mu}_m|T_m)$, $p(\sigma)$ such that each single tree T_m is a weak learner. Such specification aims at preventing the individual tree effects from unduly influential and achieving stable predictions, with automatic default specifications facilitating easy

implementation. For $p(T_m)$, the prior is specified by three aspects: (i) the probability that a node is nonterminal, (ii) the distribution on the splitting variable assignments at each interior node, and (iii) the distribution on the splitting rule assignment in each interior node, conditional on the splitting variable. For $p(\boldsymbol{\mu}_m|T_m)$ and $p(\sigma)$, conjugate normal distributions and inverse chi-square distributions are specified.

In SBART, $g(\cdot)$ associates the values of covariates with a probabilistic (instead of deterministic as in BART) path down the tree, with certain probability going left at each node. With such modification, a particular set of values of $(\boldsymbol{z}, \boldsymbol{x})$ is associated with a certain terminal node with certain probability, obtained by averaging over all possible paths. Unlike hard decision trees in BART where each terminal node is constrained to influence the regression function locally, the soft decision trees in soft BART allow each terminal node to impose a global effect on the function. This global effect of local terminal nodes enables the soft decision trees to borrow information adaptively across different covariate regions. Sparsity-inducing priors are specified to achieve a balance between sparse and non-sparse settings.

In practice, cross validation could be applied to determine the number of trees M and the hyperparameters in the Bayesian priors. Linero & Yang (2018) develop default prior specification with $M = 50$ which performs universally well in all the 10 benchmark datasets considered in the paper. In this paper, we consider $M = 50$ for BART and SBART in the simulation studies and applied examples. The BART and SBART prediction estimators of finite population mean, $\widehat{Q}_{\text{BART}}$ and $\widehat{Q}_{\text{SBART}}$, are obtained with the following steps.

Step 1 Model $p(y|\boldsymbol{z}, \boldsymbol{x})$ using BART or soft BART, $y = G(\boldsymbol{z}, \boldsymbol{x}) + \epsilon, \epsilon \sim N(0, \sigma^2)$ with corresponding Bayesian priors.

Step 2 Obtain posterior distributions of $Q(y) = \frac{1}{N}\sum_{i \in U} y_i$ using Markov chain Monte

Carlo (MCMC) simulations. Specifically, in MCMC iteration t,

1. draw $G^{(t)}, \sigma^{(t)} | y_{i\in s}, \boldsymbol{z}_{i\in U}, \boldsymbol{x}_{i\in U}$
2. compute $\tilde{\theta}_i^{(t)} = G^{(t)}(\boldsymbol{z}_i, \boldsymbol{x}_i)$ for $i \in U$
3. obtain $\widehat{Q}^{(t)}_{\text{(S)BART}} = \frac{1}{N}\left[\sum_{i\in U} \tilde{\theta}_i^{(t)} + \left(\sum_{i\in s} y_i - \sum_{i\in s} \tilde{\theta}_i^{(t)}\right)\right]$, using the observed y_i in the sample and the predicted values for the population units that are not in the sample.

Step 3 Obtain $\widehat{Q}_{\text{(S)BART}}$: point estimates using (posterior) median of $\widehat{Q}^{(t)}_{\text{(S)BART}}$ with credible intervals constructed using quantiles splitting the tails of posterior distribution equally.

In some cases, inference on subpopulation means are also of interest, which can be obtained via modification of item 3 in Step 2, restricting the average to predictions and observed outcomes in the corresponding subpopulation $\Omega \subset U$ and subsamples $s \cap \Omega$, $\widehat{Q}^{(t)}_{\Omega,\text{(S)BART}} = \frac{1}{N_\Omega}\left[\sum_{i\in\Omega} \tilde{\theta}_i^{(t)} + \left(\sum_{i\in s\cap\Omega} y_i - \sum_{i\in s\cap\Omega} \tilde{\theta}_i^{(t)}\right)\right]$.

2.2.2 BART and Soft BART Propensity Prediction

In the missing data literature, Little & An (2004) proposed including the logit-transformed response propensity score as a covariate using splines in the imputation models. This response propensity prediction method yields robust estimates of sample means when the imputation model is misspecified. Tan, Flannagan & Elliott (2019) extended the method of Little & An (2004) by using BART to fit both the imputation model and the response propensity model. They found that adding BART-estimated propensity score in the BART imputation model reduced bias and root mean square error and improves confidence interval coverage rates in the mean estimation.

Inspired by this, we extend the BART and SBART prediction with a two-step approach. First, we estimate sample inclusion propensity using a propensity model. We code the sample inclusion indicators $I = 1$ for the units in the sample and $I = 0$ for the rest of the units in the population. The estimated propensity score $\hat{\pi} = \Pr(I = 1|\boldsymbol{z}, \boldsymbol{x})$ can then be obtained via modeling $p(I|\boldsymbol{z}, \boldsymbol{x})$ using probit Bayesian additive regression trees (Chipman et al. 2010). Next, we model $p(y|\boldsymbol{z}, \boldsymbol{x}, \hat{\pi})$ by additionally including $\hat{\pi}$ as a covariate in BART or SBART model with the rest of the steps being the same as Section 2.2.1. The detailed steps of obtaining the BART propensity (BART-P) prediction estimator $\widehat{Q}_{\text{BART-P}}$ and the SBART propensity (SBART-P) prediction estimator $\widehat{Q}_{\text{SBART-P}}$ are outlined as follows.

Step 1 Model $p(I|\boldsymbol{z}, \boldsymbol{x})$ with probit BART and estimate $\hat{\pi}$ using posterior mean

Step 2 Obtain the (S)BART-P prediction estimator for finite population mean

- model $p(y|\boldsymbol{z}, \boldsymbol{x}, \hat{\pi})$ using (S)BART, $y = G(\boldsymbol{z}, \boldsymbol{x}, \hat{\pi}) + \epsilon, \epsilon \sim N(0, \sigma^2)$
- estimate $\tilde{\theta}_i^{(t)} = G^{(t)}(\boldsymbol{z}_i, \boldsymbol{x}_i, \hat{\pi}_i)$
- $\widehat{Q}^{(t)}_{\text{(S)BART-P}} = \frac{1}{N}\left[\sum_{i\in U} \tilde{\theta}_i^{(t)} + \left(\sum_{i\in s} y_i - \sum_{i\in s} \tilde{\theta}_i^{(t)}\right)\right]$
- $\widehat{Q}_{\text{(S)BART-P}}$: point estimates using (posterior) median of $\widehat{Q}^{(t)}_{\text{(S)BART-P}}$ with credible intervals constructed using quantiles splitting the tails of posterior distribution equally.

BART-P and SBART-P prediction methods are expected to be doubly robust (Long, Hsu & Li 2012): as long the mean model for either the outcome or the propensity model is correctly specified, a consistent estimator of the population mean is obtained.

3 Simulation Studies

3.1 Simulation Design

Artificial populations with size $N = 3,000$ were simulated. For each unit i in the population, a total number of p binary auxiliary variables and r continuous variables were generated. The p binary variables $\{Z_{il}\}_{l=1,\ldots p}$ were obtained with $z_{il} = I(w_{il} < u_l)$, where $\{w_{il}\} \overset{\text{i.i.d}}{\sim} \text{N}(0,1)$ and $u_l \overset{\text{i.i.d}}{\sim} \text{U}(-0.4, 0.4)$, so that $\Pr(Z_{il} = 1)$ falls in the range $(0.34, 0.66)$, $l = 1, \ldots, p$. The r continuous $\{X_{il}\}_{l=1,\ldots r}$ were generated independently from Unif$(0,1)$. Samples of size $n = 600$ were drawn from the populations with inclusion probability $\pi = \Pr(I = 1|\boldsymbol{z}, \boldsymbol{x})$ as a function of the auxiliary variables $\boldsymbol{x}$ and $\boldsymbol{z}$. We considered the following four simulation scenarios:

S1 Low-dimensional auxiliary variables ($\boldsymbol{p = 3, r = 1}$) with higher inclusion propensity at the lower tail of $\boldsymbol{X_1}$. The outcome values $\{y_i\}_{i=1,\ldots,N}$ were generated using an additive model: $y = 26.81 - z_1 - 2z_2 - 3.5z_3 - 25(x_1 - 0.75)^2 + \epsilon$, $\epsilon \sim \text{N}(0, 3^2)$, and the samples were selected with probabilities $\pi \propto \text{logit}^{-1}[-13.66 + 0.5z_1 + z_2 + 1.75z_3 + 12.5(x_1 - 0.75)^2]$. Consequently, units with x_1 falling between 0.5 and 1 were undersampled.

S2 High-dimensional auxiliary variables ($\boldsymbol{p = 30, r = 10}$) with higher inclusion propensity at the lower tail of $\boldsymbol{X_1}$. Same y and π models as S1, but add noise auxiliary variables $\{Z_l\}_{l=4,\ldots,30}$ and $\{X_l\}_{l=2,\ldots,10}$ that are not associated with y or π.

S3 High-dimensional auxiliary variables ($\boldsymbol{p = 30, r = 10}$) with the assumption A3 violated. Same as S2, but change the signs of the coefficients in the model for π to result in small π for some population units:

$\pi \propto \text{logit}^{-1}[4.01 - 0.5z_1 - z_2 - 1.75z_3 - 12.5(x_1 - 0.75)^2]$. Units with $x_1 \leq 0.25$ had lower probability of inclusion, and x_1 in the population could be outside the range in the sample.

S4 High-dimensional auxiliary variables ($p = 30, r = 10$) with more complex models and the sample has sparse data at the tails of two continuous variables that are not associated with y. The outcomes $\{y_i\}_{i=1,\ldots,N}$ were generated using

$y = 36.81 - z_1 - 2z_2 - 3.5z_3 - 10Z_1z_2 - 9(x_1 - 0.75)^2 - 16z_3(x_1 - 0.75)^2 + \epsilon, \epsilon \sim \text{N}(0, 3^2)$, with samples selected using

$\pi \propto \text{logit}^{-1}[3.27 - 0.5z_1 - z_2 - 1.75z_3 - 2z_1z_2 - 4(x_3 - 0.75)^2 - 3z_3(x_3 - 0.75)^2 - (x_5 - 0.75)^2]$. Units at the lower tails of X_3 and X_5 had lower probability of inclusion, and thus x_3 and x_5 in the population could be outside the ranges in the sample. But different from the scenario S3, X_3 and X_5 were not associated with y.

Figure 2 shows the scatterplots of y against the continuous variable X_1 in Scenarios S1–S3 and against both X_1 and $X3$ in Scenario S4. In each plot, the simulated population is overlaid with a selected sample. Population units with lower values of X_1 were more likely to be selected into samples in Scenarios S1–S2, less likely to be selected in S3, and equally likely to be selected in S4. Scenarios S3–S4 were designed to examine how the regularized prediction methods are sensitive to the assumption A3 in Section 2.2. Specifically, the values of some continuous variables (X_1 in Scenario S3 and X_3 and X_5 in Scenario S4) in the population could be outside the range of those in the sample. In Scenario S3, the out of range variable X_1 is also associated with y. In Scenario S4, the out of range variables X_3 and X_5 are not associated with y, but X_1, the variable associated with y, had the same range of values between the sample and the population.

For each scenario, 500 replicates of simulation were conducted, with point and interval

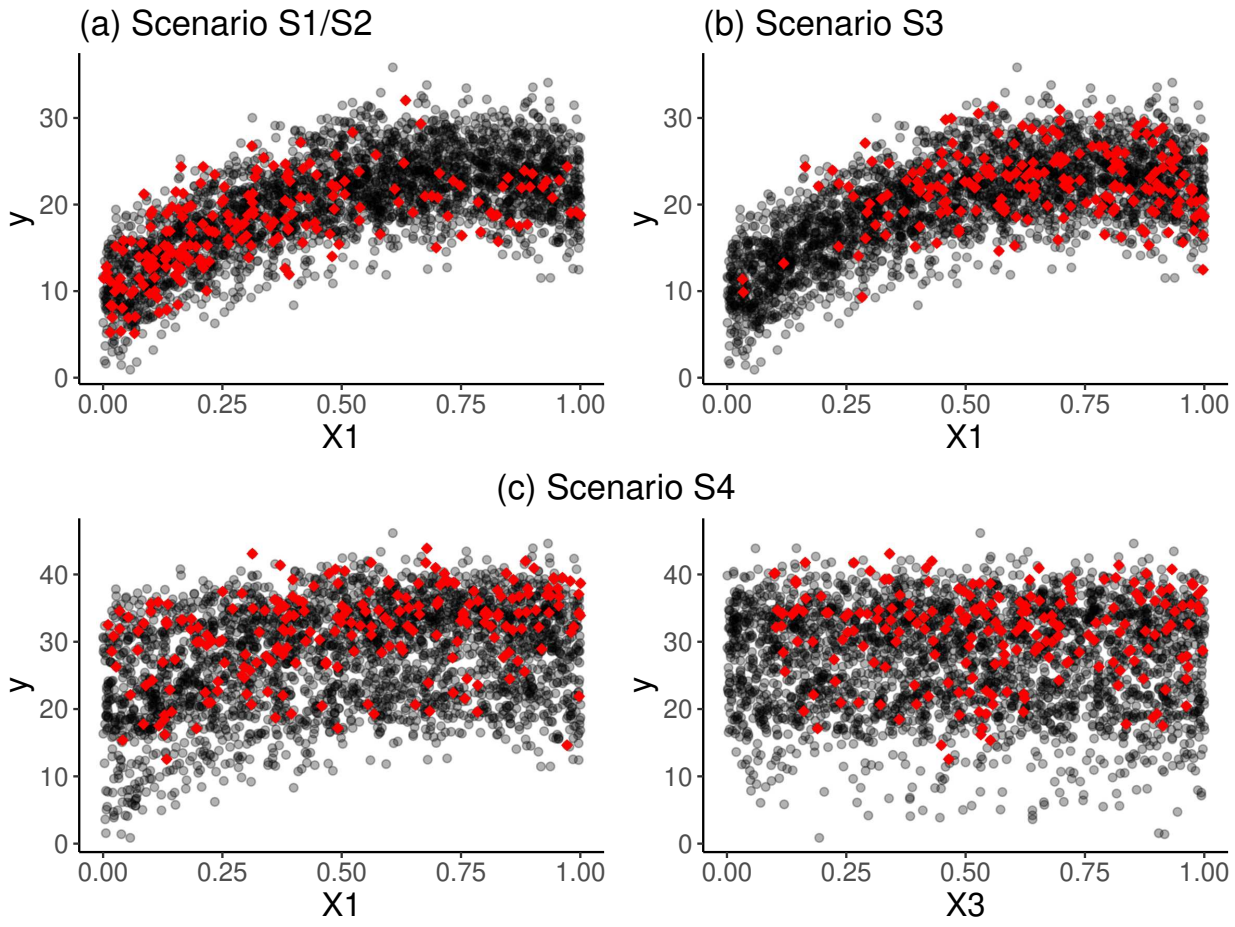

Figure 2: Scatterplots of outcomes y versus continuous auxiliary variables of units in the population (in gray dots) and a selected sample (in red diamonds) for (a) Scenario S1–S2, (b) Scenario S3, (c) Scenario S4.

estimates of finite population mean computed for each. The Bayesian additive-trees-based methods used all available auxiliary variables, as it is unknown which variables are involved in the true data generating process in practice. For scenario S1 with low-dimensional auxiliary variables, the trees-based methods were also compared to the PS and raking estimators using all four available variables with X_1 discretized using tertiles in PS and using quintiles in raking. Raw estimates were also calculated using sample means.

3.2 Simulation Results

The performance of point estimates are evaluated with empirical bias and empirical root mean square error (RMSE), summarised in Table 1. Scenarios S1–S3 share the same outcome model, the same outcome values $\{y_i\}_{i=1,\ldots,N}$ and the same ground truth for the finite population mean defined as $Q = \frac{1}{N}\sum_{i=1}^{N} y_i$. The empirical coverage rates and average widths of 80% and 95% probability intervals are visualized in Figure 3. Two horizontal dash lines denote the nominal levels of 0.8 and 0.95, respectively. The y-axis shows coverage rate and the x-axis shows average width. Estimators with coverage rates close to the nominal level lines and also small in the average interval widths are desired. The raw estimates ignoring selection bias are off the chart, leading to confidence intervals with 0% coverage rates, therefore, not shown in Figure 3.

In scenario S1, where the weighting methods are feasible, raking is less biased as well as more efficient than poststratification (PS). This is because raking maintains more information from the continuous variable X_1 by discretizing X_1 using quintiles as compared to tertiles in PS, and raking implicitly assumes an additive propensity model while PS assumes an interaction model. Both PS and raking generate confidence intervals with coverage rates lower than the nominal levels, with raking yielding shorter intervals but

Method	S1		S2		S3		S4	
	$Q = 19.88$						$Q = 27.74$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
raw	−2.99	2.99	−2.99	2.99	2.43	2.43	3.13	3.14
PS*	−0.37	0.43						
raking**	−0.16	0.22						
BART	−0.08	0.17	−0.17	0.22	0.37	0.43	0.07	0.17
BART-P	−0.06	0.18	−0.12	0.20	0.30	0.38	0.06	0.17
SBART	−0.08	0.17	−0.10	0.19	0.24	0.32	0.04	0.16
SBART-P	−0.07	0.18	−0.10	0.19	0.24	0.32	0.04	0.16

*PS is based on Z_1, Z_2, Z_3 and X_1 discretized using tertiles
**Raking is based on Z_1, Z_2, Z_3 and X_1 discretized using quintiles.
The standard errors of empirical bias from 500 simulation replicates are $< 7.5 \times 10^{-3}$ for all methods

Table 1: Simulation results: empirical bias and RMSE of various methods in estimating population means, from 500 simulation replicates, for each simulation setting

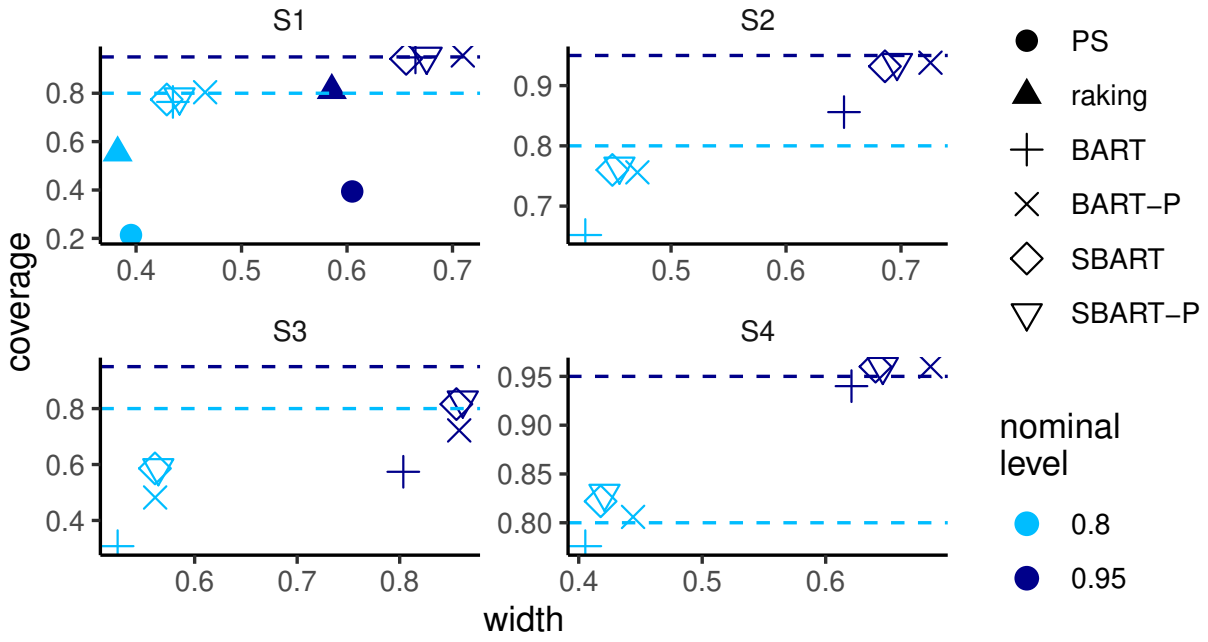

Figure 3: Simulation results - empirical coverage rates of 80% and 95% probability intervals (with the horizontal dashed lines denoting the nominal levels) against average probability interval widths, from 500 simulation replicates, for each simulation setting.

higher coverage rates. BART and SBART both outperform the weighting methods by using the continuous form of X_1, generating credible intervals with coverage rates close to the nominal levels. BART and SBART perform similarly, as all auxiliary variables are relevant in this low-dimensional setting. Including propensity score in BART and SBART leads to a small bias reduction which is offset by efficiency loss, indicated by slightly higher RMSE and slightly wider credible intervals. The coverage rates of both 80% and 95% probability intervals of the four additive-trees-based estimators are close to the nominal levels.

Scenario S2 differs from scenario S1 by adding irrelevant auxiliary variables. PS and raking are not feasible due to high-dimensionality. Units with X_1 falling between 0.5 and 1.0 have lower selection probabilities than those with X_1 in between 0 and 0.5 as shown in Figure 2(a). All the additive-trees-based methods yield much smaller bias and RMSE than the raw estimate. BART has the shortest interval widths but also lower than nominal level coverage rates for both 80% and 95% probability intervals. Including propensity score in BART reduces bias and RMSE and fixes credible interval coverage. SBART outperforms BART with lower bias, lower RMSE and close to nominal level probability interval coverage. SBART-P does not improve over SBART.

Scenario S3 differs from scenario S2 in the direction of selection bias. Moreover, because π was negatively associated with $(X_1 - 0.75)^2$, the units in the lower tail of X_1 (e.g. $X_1 < 0.25$ in Figure 2(b)) have even smaller inclusion probabilities than the units in the upper tail of X_1 in scenario S2. Consequently, there are sparse data in the lower tail of X_1 and the samples may not cover the entire range of X_1 in the population. That is, the assumption A3 is violated. In this setting, although the additive-trees-based methods yield much smaller bias and RMSE than the raw estimate, neither performs well as in Scenario S2. The empirical coverage rates for both BART and SBART are lower than the nominal levels due to bias in the estimation. By including propensity score, BART-P improves

credible interval coverage as well as bias and RMSE than BART, but does not fix the undercoverage issue. SBART yields smaller bias and RMSE and better coverage rate than BART. Similar to Scenario S2, including propensity score does not improve SBART estimate. The findings suggest that both (S)BART or (S)BART-P are sensitive to the assumption A3 in Section 2.2, when the out-of-range variables are also associated with the outcome variable y.

In scenario S4, both BART and SBART perform well with small bias and RMSE and close to nominal level coverage rate. SBART yields slightly smaller bias, smaller RMSE, and better coverage rate than BART. Including the propensity score slightly reduces bias and improves coverage rate in BART. Although there are sparse data in the lower tails of X_3 and X_5, these two X variables are not associated with y, and thus such biased selection did not yield poor performance of the additive-trees-based methods like in Scenario S3. Both BART and SBART estimators are robust to the violation of assumption A3 when the out-of-range variables are not associated with y.

3.3 Comparison of BART and SBART Predictions

We took a further investigation to compare the performance of BART and SBART in scenario S3, where the assumption A3 is violated and neither BART nor SBART performs well with BART performing worse than SBART. We consider two random samples from the population. For sample I, data are sparse at the lower tail of X_1, while, for sample II, no data are available at the lower tail, $X_1 < 0.2$. The top panels I(a) and II(a) of Figure 4 shows the population in gray dots, with sample I and II in red, respectively. For a closer examination of the data at the lower tail of X_1, we restrict to a subset with $Z_2 = Z_3 = 0$ and focus on the lower tail with $X_1 < 0.3$. The middle panels I(b) and II(b) show the

population and corresponding sample data in this restricted subgroup. Finally, the bottom panels I(c) and II(c) plot the population units of y in this subgroup (gray points) overlaid by the posterior means of the location parameters, $G(\boldsymbol{Z}, \boldsymbol{X})$, of each population units estimated using BART and SBART as shown using red pluses and blue crosses, respectively. Panel I(c) shows that both BART and SBART fit the data well within the region $X_1 > 0.2$ where sample data are available. However, the SBART fit the data much better than BART when $X_1 < 0.2$ where sparse data are available. The estimated posterior means of the location parameters based on SBART are also less noisy, due to the sparsity-induced priors that tend to exclude the noise auxiliary variables in model fitting. Panel II(c) shows that both models fail to produce valid predictions in the region $X_1 < 0.2$ where there is no sample data available. In the simulation study, about 5% of the 500 simulated samples do not include units with $X_1 < 0.1$ and one third include fewer than 10 units with $X_1 < 0.2$. SBART outperforms BART in all simulation scenarios considered and is, therefore, recommended.

4 Applied Examples

We demonstrate the application of the proposed methods using real data from two different studies. The first application example deals with a mental health survey assessing psychiatric disorders among the Ohio Army National Guard (OHARNG) service members. The second application is in a clinic setting where it is of interest to generalize inference on COVID-19 patients when clinical outcomes are only available in a subset of patients.

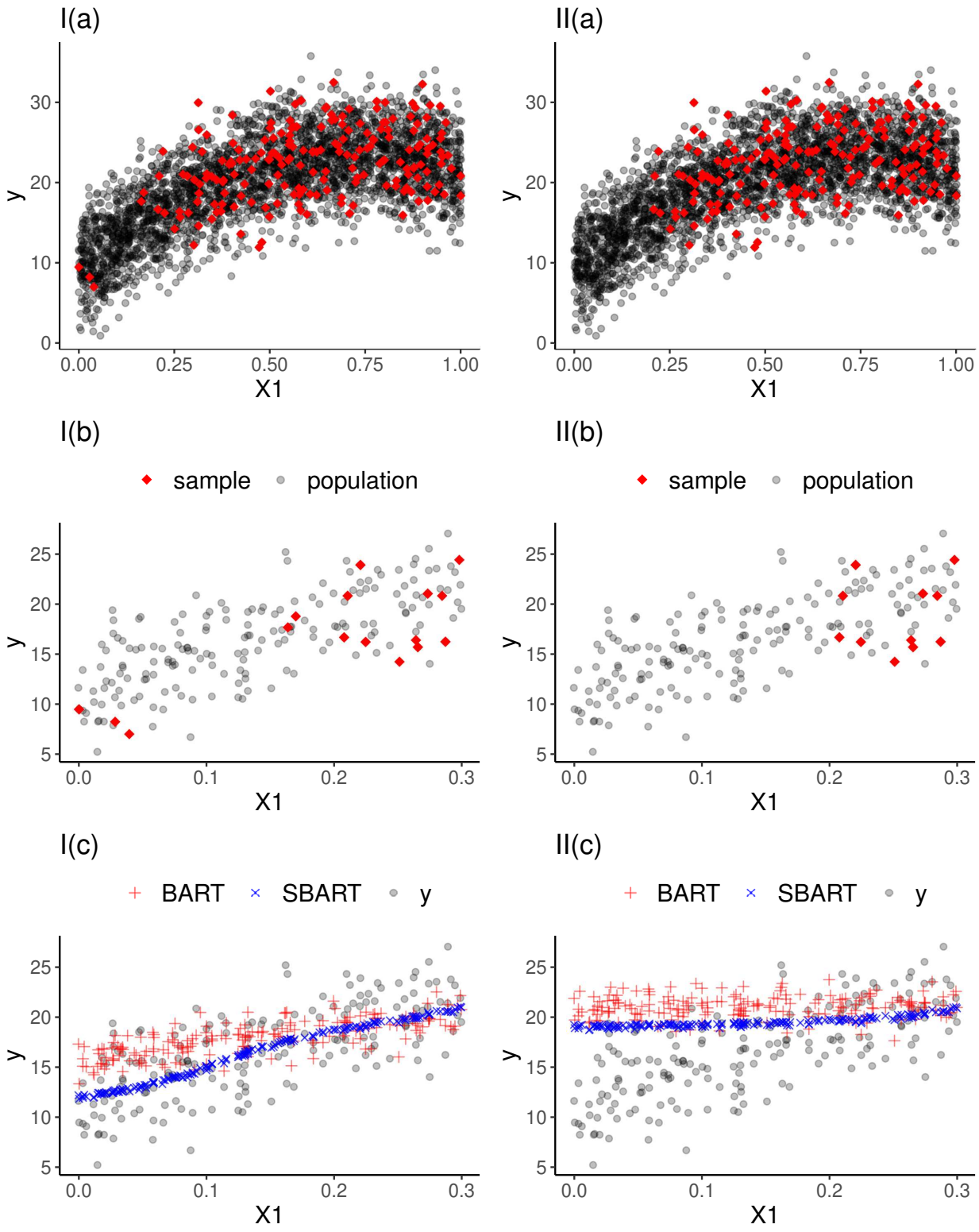

Figure 4: Two selected samples I and II from the population in Scenario S3: (a) Scatterplots of y versus X_1 with the population in gray dots and a selected sample in red diamonds; (b) Scatterplots of y versus X_1, restricted to $Z_2 = Z_3 = 0$ and $X_1 < 0.3$ (c) Scatterplots of y versus X_1 in the subpopulation, overlapped with posterior means of $G(\boldsymbol{Z}, \boldsymbol{X})$ estimated from the BART and SBART models based on the whole sample.

4.1 Ohio Army National Guard Survey of Mental Health

The Ohio Army National Guard (OHARNG) Mental Health Initiative is a population-based observational survey study for estimating the prevalence and identifying correlates of mental illness and health service use among the OHARNG service members. The study population of the baseline survey is defined as all $N = 12570$ soldiers who served in the OHARNG between June 2008 and February 2009. A survey sample with $n = 2562$ service members was selected. The study was designed as a probability survey, but the sample was subject to serious selection bias due to nonresponse and undercoverage of sampling frame. In this analysis, we are interested in estimating the mean trauma score among the OHARNG service members using the selected sample. Auxiliary information is available at individual level for the entire study population, including age (17–24 yr, 25–34 yr, 35+ yr), sex (male, female), race (white, black, other), rank (enlisted, officer), marital status (married, non-married), and years of service (in years). We apply the proposed trees-based methods to correct the discrepancy between the sample and the population. Although the indicator variable I is not observed for the population units, we matched the sample and the population units using the five categorical and one continuous auxiliary variables. Once the indicator variable I was created, for BART-P and SBART-P, the propensity models were built using probit BART.

Before modeling, we followed standard practice and applied the $\log(y + 1)$ transformation to trauma scores. Distributions of the only continuous variable, years of service, and the estimated propensity score $\hat{\pi}$ in the sample and population were checked to avoid prediction failure due to sparse data at the tails (see Figure S1 in the Supplementary Materials). After fitting the models, we performed posterior predictive checks (Gelman et al. 2014, chapter 6) based on the following test quantities: (a) $T_1(\boldsymbol{y}) = \bar{y}$, (b) $T_2(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, and (c) $T_3(\boldsymbol{y}, G, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\theta_i}{\sigma}\right)^2$, where $\theta_i = G(\boldsymbol{z}_i, \boldsymbol{x}_i)$.

The test quantities catch different aspects of the data, with $T_1(\cdot)$ and $T_2(\cdot)$ measuring the location and variability of the survey outcome while $T_3(\cdot)$ measuring the discrepancy between the survey outcome and fitted distribution. In each MCMC iteration t, the realized test quantities $T_i(\boldsymbol{y}, G^{(t)}, \sigma^{(t)})$ under the observed data and predictive test quantities $T_i(\tilde{\boldsymbol{y}}^{(t)}, G^{(t)}, \sigma^{(t)})$ under the simulated data were computed and compared, with $\tilde{\boldsymbol{y}}^{(t)}$ drawn from the posterior predictive distribution. For each quantity $T_i(\cdot)$, a Bayesian posterior predictive p-value can also be computed, which is defined as the probability that the predictive test quantity is greater than the realized test quantity, evaluated over the posterior distribution. The Bayesian p-value measures the discrepancy between the observed data and the posterior predictive distribution in the aspect characterized by $T(\cdot)$. A Bayesian p-value close to 0.5 indicates a predictive distribution close to the observed data, while a Bayesian p-value near 0 or 1 correspond to data in the tails of the posterior distribution, thus representing a misfit to the model. Figure S2 in the Supplementary Materials shows the posterior predictive graphics checking and corresponding p-values for SBART. All four Bayesian methods, BART, BART-P, SBART, and SBART-P, yielded fitted models with posterior predictive p-values close to 0.5, indicating predictive distributions that covered the observed data.

We compare the results of proposed Bayesian methods with the raw estimates in estimating the mean trauma score on the log scale, with point estimates and 95% probability intervals visualized in Figure 5(a). The Bayesian estimates of mean trauma score are lower than the raw unadjusted estimates. BART and SBART yield similar results, as this is a low-dimensional setting with one continuous auxiliary variable and the benefit of soft decision trees is not so obvious. Including propensity scores do not lead to much change in the estimates. We recommend reporting the estimates using SBART for this problem.

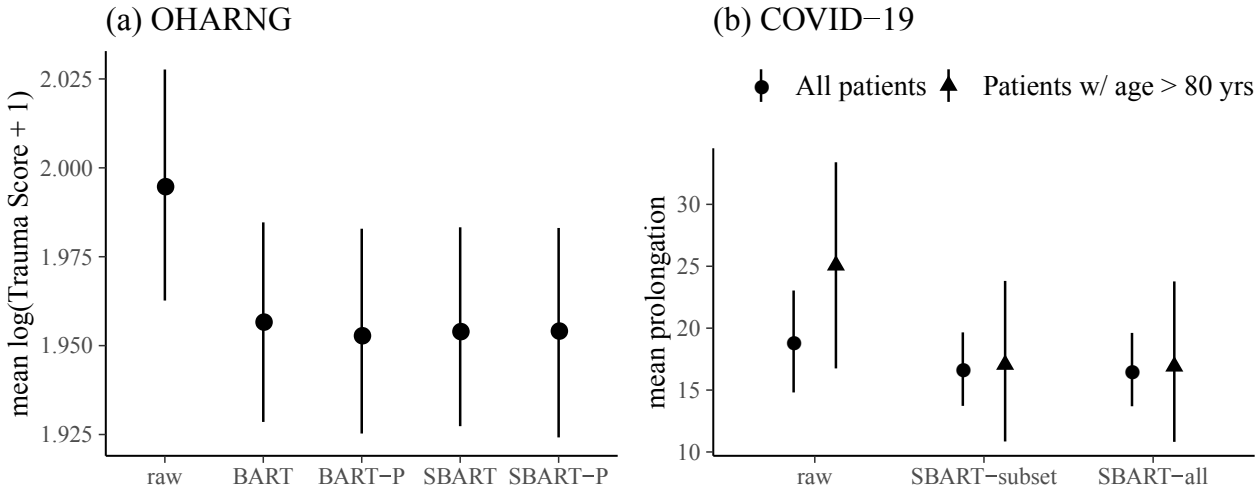

Figure 5: (a) Point estimates and 95% probability intervals of mean log(trauma score + 1) among soldiers who served in the OHARNG between June 2008 and February 2009 (b) Point estimates and 95% probability intervals of mean prolongation among all patients and patients with age $\geq$ 80 years old, comparing raw sample means, SBART with baseline QTc and treatment (SBART-subset), and SBART with all auxiliary variables (SBART-all).

4.2 New York City COVID-19 Study

The first positive case of COVID-19 was confirmed in New York City on March 1, 2020. By May, the city had more cases than any country other than the United States. The urgent need for therapeutic agents has resulted in repurposing and redeployment of experimental agents. Hydroxychloroquine, combined with azithromycin, was administered to patients with COVID-19 without robust evidence supporting its use, with the U.S. Food and Drug Administration (FDA) issuing an emergency use authorization (EUA) to allow doctors to begin treating patients with hydroxychloroquine in hospitalized settings outside clinical trials on March 28, 2020. This EUA was later revoked as of June 15, 2020, following a randomized clinical trial in hospitalized patients that showed no benefits and reports of serious heart rhythm problems along with other safety issues. Both hydroxychloroquine and azithromycin are characterized as definite QTc prolongers that increase risk of sudden cardiac deaths.

Between March 1 through May 1st, 2020, 470 patients were admitted to Columbia University Irving Medical Center and treated with hydroxychloroquine (H+) or hydroxychloroquine combined with azithromycin (A+H+) (Rubin et al. 2021+). We are interested in estimating the mean QTc prolongation, defined as difference in QTc measures between day 2 and day 0, of all the 470 COVID-19 patients who received H+ or A+H+ treatments. Although all patients have baseline ECG measurements of QTc, only 244 of them have ECG QTc measurements on Day 2 of medication. These 244 patients formed a non-probability sample of the target population of all 470 patients. To improve the estimation for the population mean of QTc prolongation using the 244 patients, we also obtained the data of all 470 patients on their demographic characteristics and relevant biomarkers from electronic medical records (EMRs). Patients in the sample are linked to the population data in the EMRs and the propensity model was built using probit BART.

Exploratory analysis indicates a strong negative association between prolongation and baseline QTc measurement, that patients with higher baseline QTc are less likely to have ECG QTc measurement on Day 2, and similar ranges for values of baseline QTc measurement and the estimated propensity score between the sample and population, demonstrated in Figure S3 in the Supplementary Materials. Other auxiliary variables include treatment (H+, A+H+), demographic characteristics, including gender, age (in years), race (white, black, other), BMI (log scale), along with 7 biomarkers. We used the recommended method SBART for two estimands of interest: (i) the mean QTc prolongation among all 470 patients, and (ii) the mean QTc prolongation among the 87 (out of 470) patients who were over 80 years old. We compared two SBART models, with the first model only including baseline QTc and treatment (SBART-subset) and the second model including all covariates (SBART-all). SBART-P performs similarly to SBART and is not presented here.

As is visualized in Figure 5(b), SBART yields lower estimates of mean QTc prolongation compared to the raw estimates ignoring selection bias, for both estimands of interest. For estimand (i), including baseline QTc and treatment in SBART leads to obvious drop in the mean prolongation estimates from a raw estimate of 18.9 (95% CI : $14.8, 23.0$) milliseconds to 16.7 (95% CI : $13.7, 19.7$) milliseconds. Additionally adding other auxiliary variables does not lead to further obvious change in the estimates.

For estimand (ii), estimates can be readily obtained by restricting the calculation to predicted and observed prolongation for patients over 80 years old, without additional computation for model fitting. Similar to the estimate among all patients, SBART yields smaller estimates and shorter probability intervals than the raw estimator. Although the mean prolongation estimate is higher among this subgroup of patients than all patients using the raw estimator, the estimates are similar using SBART.

5 Discussion

We consider generalization of inference on a descriptive estimand from a non-random sample to a target population in data-rich settings where high dimensional auxiliary information is available in both the sample and population, with survey inference being a special case. Existing methods such as poststratification, raking, and MRP are challenging or infeasible due to high dimensionality, and the need to discretize continuous auxiliary variables before applying such methods leads to loss of information. To address such issues, we propose a regularized prediction approach by modeling the conditional distribution of the outcomes given the high-dimensional auxiliary variables using Bayesian machine learning techniques. In this paper, we specifically consider BART and soft BART which

handles both discrete and continuous auxiliary variables as well as potential interactions. Besides the auxiliary variables, we also consider modified methods that estimates the propensity score for a unit to be included in the sample and also include the estimated propensity score as a covariate in the BART and soft BART model.

Artificial data simulation studies demonstrate that the Bayesian additive-trees-based methods outperform poststratification (PS) and raking in low-dimensional settings where PS and raking are feasible, as the regularized additive trees better use information in the continuous auxiliary variables and avoid model overspecification. The Bayesian additive-trees-based methods also yield valid inference in high-dimensional settings when PS and raking are not feasible, as long as selection bias does not result in sparse data points at the tails of some continuous auxiliary variables that are predictive of the survey outcome variable y. In high-dimensional setting with sparse signals, SBART, with soft decision trees and sparsity-inducing priors, is less biased and more efficient than BART. In challenging settings where the additive-trees-based methods underperform, including propensity score in BART could reduce bias and improve credible interval coverage while such benefit is not obvious for SBART. Therefore, the soft BART prediction method is recommended for generalization of inference with high-dimensional auxiliary variables. The soft BART better uses information in the continuous auxiliary variables and more effectively regularize the effect of irrelevant noise auxiliary variable. As is demonstrated in the OHARNG mental health study and the COVID-19 study, the proposed methods could be applied in both surveys and more general settings, with estimands being overall population as well as subpopulation quantities.

The Bayesian additive-trees-based methods need to be used with caution. Both BART and soft BART prediction fail when selection bias results in no data points at the tails of the continuous covariates that are associated with the outcomes. Such prediction failure cannot

be fixed via robust methods involving propensity scores. Therefore, for important continuous variables associated with the outcomes, the range and distribution in the sample and population need to be checked before using the methods. In some cases, transformation on such auxiliary variables could be applied to improve prediction at the tails. However, as we showed in the simulation Scenario S4, the proposed methods can still work well when the sample does not contain data points at the tails of the continuous variables that are not associated with the outcomes even those these variables are associated with selection.

We performed a series of sensitivity analyses under the simulation Scenario S2 (see Table S1 in the Supplementary Materials). First, we modified scenario S2 to allow all the auxiliary variables to be correlated. All four additive-trees-based methods perform well, with SBART yielding smaller bias and closer to the nominal level coverage rate than BART. Second, we assessed whether our proposed methods are robust to the positivity assumption, allowing some population units to have zero chance of being selected. The analysis suggests that the positivity assumption is not necessary for the proposed methods as long as the three required assumptions in Section 2.2 are satisfied. Third, we evaluated the robust property of the (S)BART-P approach by excluding the important continuous variable X_1 from the additive-trees-based predictive models for y. The analysis shows that now (S)BART performs poorly due to omitting an important predictor of y, but the (S)BART-P are robust to the model misspecification and perform well. Last, we examined the impact of using estimated propensity scores by recomputing the (S)BART-P estimators in scenario S2 by replacing the estimated propensity score with the true propensity score in the data simulation. The new estimators have similar bias, RMSE, and coverage rate of 95% probability interval as the (S)BART-P in Scenario S2.

Linking sample and population units is not easy in a high-dimensional setting if the inclusion indicator I is not available. However, data or record linkage techniques have been

well developed and are widely available nowadays (Bohensky et al. 2010). Although data linkage may be associated with errors if the sample and population data do not consistently record information from the same cases (e.g. rounding errors for continuous variables), the impact of such errors on our proposed methods is rather limited for two reasons. First, our goal is to use data linkage for creating the sample inclusion indicator I and estimating the propensity score. As long as the sample units have a close match to the population units (i.e., the values of the auxiliary variables in the sample units are close to the values of the linked population units), the estimates of propensity score would be close enough to the estimates without linkage errors. Second, the estimated propensity score is only used as a predictor in the predictive models of our proposed methods, so minor measurement error would have limited impact on our estimates of population quantities, especially given that the SBART approach with and without including the propensity score lead to similar estimates of population quantities.

The paper assumes that unit-level auxiliary information is available for all the units in the target population. This is the case for our two application examples. In the era of data science, high-dimensional population data become more widely available and accessible. Such data can include administrative records and electronic medical records. Whether an external dataset can be treated as the population data used in our methods largely depends on the target population of interest, and thus the population data should be chosen carefully. The proposed methods can also be naturally extended to handle the setting where unit-level population data are not available. When high-dimensional auxiliary variables are available in a reference probability survey, a two-step approach can be used (Rafei, Flannagan, West & Elliott 2021). The outcome variables can be first predicted based on the SBART model and the complex survey design information can then be applied to the predicted values of the outcome variables in the reference survey to obtain population inference. The resulting variance estimation needs to account for not

only the uncertainty in the predictive models but also the sampling error in the reference probability survey. Alternatively, synthetic populations can be generated using the weights in the reference survey when the population is not extremely large (Zangeneh & Little 2015, Dong et al. 2014). The problem of combining non-random samples with a reference probability survey is also discussed by other researchers.

The key assumption for the validity of prediction inference is that the sampling mechanism is ignorable. That is, the sample inclusion indicator I and the outcome y are independent given all the available auxiliary variables $(\boldsymbol{Z}, \boldsymbol{X})$. The ignorable assumption is reasonable when there include a large number of auxiliary variables, hence the need for statistical methods that can adjust for a high-dimensional $(\boldsymbol{Z}, \boldsymbol{X})$. However, when the population data do not contain enough auxiliary information, the sampling mechanism can be still informative even after adjusting for all available auxiliary variables. If that happened, simply assuming the non-informative sampling mechanism may lead to misleading results.

The advantage of BART in estimating propensity score and handling missing data and causal inference problems has been elucidated in other papers (Tan et al. 2019, Rafei et al. 2020, Kern et al. 2016, Wendling et al. 2018). However, Soft BART (SBART) is less known or used in the survey literature. SBART outperforms BART in all the simulation scenarios and is recommend for use in the prediction inference. When the same list of auxiliary variables are available in fitting additive-trees-based models for both y and I, SBART-P does not improve over SBART. However, if some key variables predictive of y are omitted from the pool of candidate variables for fitting the model for y, the SBART-P can offer protection from model misspecification in SBART. Finally, although we consider BART and SBART in this paper, the regularized prediction approach is general and any Bayesian machine learning techniques that achieve valid predictions could be applied.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. & Tourangeau, R. (2013), 'Summary report of the AAPOR task force on non-probability sampling', *Journal of Survey Statistics and Methodology* **1**(2), 90–143.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I. & Brand, C. A. (2010), 'Data linkage: A powerful research tool with potential problems', *BMC Health Services Research* **10**(1), 1–7.

Chipman, H. A., George, E. I. & McCulloch, R. E. (2010), 'BART: Bayesian additive regression trees', *Annals of Applied Statistics* **4**(1), 266–298.

Deming, W. E. & Stephan, F. F. (1940), 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics* **11**(4), 427–444.

Dong, Q., Elliott, M. R. & Raghunathan, T. E. (2014), 'A nonparametric method to generate synthetic populations to adjust for complex sampling design features', *Survey Methodology* **40**(1), 29.

Elliott, M. R. & Valliant, R. (2017), 'Inference for nonprobability samples', *Statistical Science* **32**(2), 249–264.

Gelman, A. (2007), 'Struggles with survey weighting and regression modeling (with discussion)', *Statistical Science* **22**(2), 153–164.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014), *Bayesian Data Analysis, third edition*, CRC Press.

Gelman, A. & Little, T. C. (1997), 'Poststratification into many categories using hierarchical logistic regression', *Survey Methodology* **23**(2), 127–135.

Hahn, P. R., Murray, J. S. & Carvalho, C. M. (2020), 'Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)', *Bayesian Analysis* **15**(3), 965–1056.

Hill, J. L. (2011), 'Bayesian nonparametric modeling for causal inference', *Journal of Computational and Graphical Statistics* **20**(1), 217–240.

Keiding, N. & Louis, T. A. (2016), 'Perils and potentials of self-selected entry to epidemiological studies and surveys', *Journal of the Royal Statistical Society: Series A* **179**(2), 319–376.

Kern, H. L., Stuart, E. A., Hill, J. & Green, D. P. (2016), 'Assessing methods for generalizing experimental impact estimates to target populations', *Journal of Research on Educational Effectiveness* **9**(1), 103–127.

Kim, J. K., Park, S., Chen, Y. & Wu, C. (2021), 'Combining non-probability and probability survey samples through mass imputation', *Journal of the Royal Statistical Society: Series A* **184**(3), 941–963.

Linero, A. R. & Yang, Y. (2018), 'Bayesian regression tree ensembles that adapt to smoothness and sparsity', *Journal of the Royal Statistical Society: Series B* **80**(5), 1087–1110.

Little, R. & An, H. (2004), 'Robust likelihood-based analysis of multivariate data with missing values', *Statistica Sinica* **14**(3), 949–968.

Long, Q., Hsu, C.-H. & Li, Y. (2012), 'Doubly robust nonparametric multiple imputation for ignorable missing data', *Statistica Sinica* **22**, 149.

Rafei, A., Flannagan, C. A. C., West, B. T. & Elliott, M. R. (2021), 'Robust Bayesian inference for big data: Combining sensor-based records with traditional survey data', *arXiv preprint arXiv:2101.07456* .

Rafei, A., Flannagan, C. A. & Elliott, M. R. (2020), 'Big data for finite population inference: applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees', *Journal of Survey Statistics and Methodology* **8**(1), 148–180.

Rothwell, P. M. (2005), 'External validity of randomised controlled trials: "To whom do the results of this trial apply?"', *Lancet* **365**(9453), 82–93.

Rubin, G. A., Desai, A. D., Chai, Z., Wang, A., Chen, Q., Wang, A. S., Kemal, C., Baksh, H., Biviano, A., Dizon, J. M., Yarmohammadi, H., Ehlert, F., Saluja, D., Rubin, D. A., Morrow, J. P., Avula, U. M. R., Berman, J. P., Kushnir, A., Abrams, M. P., Hennessey, J. A., Elias, P., Poterucha, T. J., Uriel, N., Kubin, C. J., LaSota, E., Zucker, J., Sobieszczyk, M. E., Schwartz, A., Garan, H., Waase, M. P., & Wan, E. Y. (2021+), 'COVID-19 infection is associated with QTc prolongation', *JAMA Network Open* .

Smith, T. (1983), 'On the validity of inferences from non-random samples', *Journal of the Royal Statistical Society: Series A* **146**(4), 394–403.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. & Leaf, P. J. (2011), 'The use of propensity scores to assess the generalizability of results from randomized trials', *Journal of the Royal Statistical Society: Series A* **174**(2), 369–386.

Tan, Y. V., Flannagan, C. A. & Elliott, M. R. (2019), 'Robust-squared imputation models using BART', *Journal of Survey Statistics and Methodology* **7**(4), 465–497.

Valliant, R. (1993), 'Poststratification and conditional variance estimation', *Journal of the American Statistical Association* **88**(421), 89–96.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015), 'Forecasting elections with non-representative polls', *International Journal of Forecasting* **31**(3), 980–991.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. & Gallego, B. (2018),

'Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases', *Statistics in Medicine* **37**(23), 3309–3324.

Zangeneh, S. Z. & Little, R. J. (2015), 'Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample', *Journal of Survey Statistics and Methodology* **3**(2), 162–192.