Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(4) 1079–1093

DOI: 10.1177/0962280217746191
journals.sagepub.com/home/smm
SAGE

Decision making and uncertainty quantification for individualized treatments using Bayesian Additive Regression Trees

Brent R Logan,[1] Rodney Sparapani,[1] Robert E McCulloch[2] and Purushottam W Laud[1]

Abstract

Individualized treatment rules can improve health outcomes by recognizing that patients may respond differently to treatment and assigning therapy with the most desirable predicted outcome for each individual. Flexible and efficient prediction models are desired as a basis for such individualized treatment rules to handle potentially complex interactions between patient factors and treatment. Modern Bayesian semiparametric and nonparametric regression models provide an attractive avenue in this regard as these allow natural posterior uncertainty quantification of patient specific treatment decisions as well as the population wide value of the prediction-based individualized treatment rule. In addition, via the use of such models, inference is also available for the value of the optimal individualized treatment rules. We propose such an approach and implement it using Bayesian Additive Regression Trees as this model has been shown to perform well in fitting nonparametric regression functions to continuous and binary responses, even with many covariates. It is also computationally efficient for use in practice. With Bayesian Additive Regression Trees, we investigate a treatment strategy which utilizes individualized predictions of patient outcomes from Bayesian Additive Regression Trees models. Posterior distributions of patient outcomes under each treatment are used to assign the treatment that maximizes the expected posterior utility. We also describe how to approximate such a treatment policy with a clinically interpretable individualized treatment rule, and quantify its expected outcome. The proposed method performs very well in extensive simulation studies in comparison with several existing methods. We illustrate the usage of the proposed method to identify an individualized choice of conditioning regimen for patients undergoing hematopoietic cell transplantation and quantify the value of this method of choice in relation to the optimal individualized treatment rule as well as non-individualized treatment strategies.

Keywords

Individualized treatment rules, prediction models, BART, boosting, random forests, outcome weighted learning, subgroup analysis, optimal ITR, value function estimation

1 Introduction

There is increasing recognition in clinical trials that patients are heterogeneous and may respond differently to treatment. A major goal of precision medicine is to identify which patients respond best to which treatments and tailor the treatment strategy to the individual patient. This personalization of treatment based on patient clinical features, biomarkers, and genetic information is formalized as an individualized treatment rule (ITR) by Qian and Murphy.[1] Individualized treatment rules extend classical subgroup analysis, in which pre-specified subgroups of the population are assessed for differential treatment effects, to the point where the treatment benefit for each individual is used to determine a treatment assignment rule that is, in some sense, optimal.

[1]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA
[2]School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA

Corresponding author:
Brent R Logan, Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Rd., PO Box 26509, Milwaukee, WI 53226-0509, USA.
Email: blogan@mcw.edu

Several strategies for identifying treatment rules have been proposed in the literature. Many of these utilize a model for the conditional mean function of the outcome given treatment and covariates, and optimize it over available treatments to define an ITR. Qian and Murphy[1] show that good prediction accuracy of the outcome model is sufficient in order to ensure good performance of the associated ITR. As a result a number of strategies for flexible prediction models have been proposed, including a large linear approximation space with penalization to avoid overfitting,[1,2] generalized additive models,[3] boosting,[4] random forests,[5] support vector regression,[6] kernel ridge regression,[7] and tree-based methods or recursive partitioning.[8–13] An alternative strategy is to directly optimize an estimator of the expected outcome of a treatment rule over a class of potential rules.[14–17] This has the advantage of not requiring an accurate prediction model, but does require specification of the class of treatment rules allowed. While each method uses its own "optimal" choice, in this article, we reserve the phrase "optimal ITR" for the ITR defined by Qian and Murphy[1] in their equation (1), and also stated in our equation (1).

The Bayesian framework leads to natural quantification of uncertainty that allows construction of credible and prediction intervals. A major distinguishing feature as described later is that our method provides direct inference on the value of the optimal ITR while it is not clear how this can be done with other existing methods. The Bayesian nonparametric regression method we have implemented here is Bayesian Additive Regression Trees (BART)[18] to model the conditional mean function of the outcome and to inform identification of an ITR. BART has been shown to be efficient and flexible with performance comparable to or better than its non-Bayesian competitors such as boosting, lasso, MARS, neural nets, and random forests.[18] Furthermore, our simulations described later in the article show that BART ITR's performance was better than or comparable to other existing methods in this setting of identifying an ITR. Because of its tree-based structure, BART can effectively address interactions among variables, which is important in this context for identification of treatment interactions leading to differential treatment recommendations. In addition, recent modifications to BART have been proposed that maintain excellent out-of-sample predictive performance even when a large number of additional irrelevant regressors are added.[19]

We present our work in the following sequence. In section 2, we describe the notation for individualized treatment rules. Following that, we review the BART methodology briefly in section 3. Section 4 describes our proposed BART-based ITR and addresses estimation of its value function as well as that of the optimal ITR. In section 5, we show excellent performance of the proposed BART ITR in a benchmark simulation comparison to existing methods for ITRs. In section 6, we conduct additional simulations to examine the operating characteristics of the BART prediction model and the estimation of the value function. In section 7, we discuss ways to approximate the BART ITR to get an interpretable clinical rule for treatment assignment, similar to identification of subgroups of patients who would benefit from assignment to particular treatments. Section 8 illustrates the usage of the proposed method on a medical application in hematopoietic cell transplantation. The article ends with a discussion of our contribution as well as of some planned future developments.

2 Individualized treatment rules

Let Y be the outcome of interest, with higher values of Y being more desirable. We focus on a binary outcome in this article, where $Y=1$ indicates a favorable outcome and $Y=0$ its complement, but the proposed method could easily be used with a continuous outcome as well. Let $\mathcal{A}$ be the treatment space with $\mathcal{A}=\{-1,1\}$ and let $X=(x_1,\ldots,x_p)$ be the vector of patient characteristics being used to personalize treatment, with population space Ω_X defined so that $p(A=1|X=x)$ is bounded away from 0 and 1, i.e., $\Omega_X=\{X: p(A=a|X=x)\in(0,1), \quad \forall a\in\mathcal{A}\}$. We assume observations represent a random sample $(Y,\ A,\ X)$ either from a randomized trial or observational study. For an observational study, we make the usual assumptions to allow causal inference that treatment assignment is strongly ignorable, i.e., treatment A is independent of the potential outcomes $Y|A=1$ and $Y|A=0$ given X. Note that BART models have been proposed for use in observational studies.[20]

A treatment rule $g(X)$ is a mapping from the covariate space Ω_x to the treatment space $\mathcal{A}$ so that patients with covariate value X are assigned to treatment $g(X)$. Let P be the distribution of $(Y,\ A,\ X)$ and $E(Y)$ be the expectation with respect to P. Let P^g be the distribution of $(Y,\ A,\ X)$ given that $A=g(X)$ and let $E^g(Y)$ be the expectation with respect to P^g. The value function $V(g)$ of a treatment rule $g(X)$ is the expected outcome associated with that treatment rule, i.e., $V(g)=E^g(Y)$. An optimal ITR g_0 is a treatment rule that optimizes the value function

$$g_0\in\arg\max_{g\in G}V(g)$$

where G is a collection of possible treatment rules. Since the value function can be written as

$$V(g) = E[E(Y|X, A = g(X))],$$

Qian and Murphy[1] show that the optimal ITR satisfies

$$g_0(X) \in \arg\max_{a \in \mathcal{A}} E(Y|X, A = a) \text{ a.s.} \tag{1}$$

so that one possible solution is to assign to each patient the treatment which has the higher conditional expectation given their covariate vector X. In practice, this strategy requires modeling of the conditional mean function and use of the estimated conditional mean function in the above expression to determine treatment assignment. Qian and Murphy[1] showed that if the prediction error of such a model is small, then the reduction in value of the associated ITR g compared to the optimal ITR g_0 is also small, pointing to the need for a flexible and accurate prediction model for the conditional mean function.

3 BART methodology

As BART is based on an ensemble of regression tree models, we begin with a simple example of a regression tree model. We then describe how BART uses an ensemble of regression tree models for a numeric outcome. Finally, we describe how the BART model for a numeric outcome is augmented to model a binary outcome.

Suppose y_i represents the (numeric) outcome for individual i, and $\boldsymbol{x}_i$ is a vector of covariates with the regression relationship $y_i = h(\boldsymbol{x}_i; T, M) + \epsilon_i$. Notationally, $h(\boldsymbol{x}_i; T, M)$ is a binary tree function with components T and M that can be described as follows. T denotes the tree structure consisting of two sets of nodes, interior and terminal, and a branch decision rule at each interior node which typically is a binary split based on a single component of the covariate vector. An example is shown in Figure 1 wherein interior nodes appear as circles, and terminal nodes as rectangles. The second tree component $M = \{\mu_1, \ldots, \mu_b\}$ is made up of the function values at the terminal nodes.

BART employs an ensemble of such trees in an additive fashion, i.e., it is the sum of m trees where m is typically large such as 200. The model can be represented as

$$\left.\begin{aligned} y_i &= f(\boldsymbol{x}_i) + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2) \\ f(\boldsymbol{x}_i) &= \sum_{j=1}^{m} h(\boldsymbol{x}_i; T_j, M_j) \end{aligned}\right\} \tag{2}$$

To proceed with the Bayesian specification, we need a prior for f. Notationally, we use

$$f \sim BART \tag{3}$$

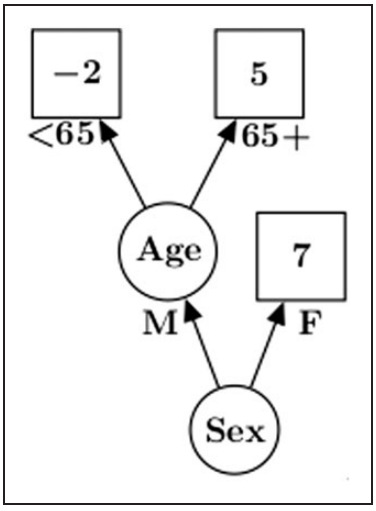

Figure 1. An example of a single tree with branch decision rules and terminal nodes.

and describe it as made up of two components: a prior on the complexity of each tree, T_j, and a prior on its terminal nodes, $M_j|T_j$. Using the Smith-Gelfand bracket notation[21] as a shorthand notation for writing a probability density or conditional density, we write $[f] = \prod_j [T_j][M_j|T_j]$. Following Chipman et al.,[18] we partition $[T_j]$ into three components: the tree structure or process by which we build a tree and create interior nodes, the choice of a covariate given an interior node and the choice of decision rule given a covariate for an interior node. The probability of a node being interior is defined by describing the probabilistic process by which a tree is grown. We start with a tree which is a single node and then recursively let a node have children (so that it is not a terminal node) with probability $\alpha(1+d)^{-\gamma}$ where d represents the branch depth, $\alpha \in (0,1)$ and $\gamma \geq 0$. We assume that the choice of a covariate given an interior node and the choice of decision rule branching value given a covariate for an interior node are both uniform. We then use the prior $[M_j|T_j] = \prod_{\ell=1}^{b_j} [\mu_{j\ell}]$ where b_j is the number of terminal nodes for tree j and $\mu_{j\ell} \sim \mathrm{N}(0, \tau^2/m)$ on the values of the terminal nodes. This gives $f(\boldsymbol{x}) \sim N(0, \tau^2)$ for any $\boldsymbol{x}$ since the value $f(\boldsymbol{x})$ will be the sum of m independent $N(0, \tau^2/m)$. Along with centering of the outcome, these default prior mean and variance are specified such that each tree is a "weak learner" playing only a small part in the ensemble; more details on this can be found in literature.[18]

To apply the BART model to a binary outcome, we use a probit transformation

$$p(Y = 1|\boldsymbol{x}) \equiv p(\boldsymbol{x}) = \Phi(\mu_0 + f(\boldsymbol{x}))$$

where Φ is the standard normal cumulative distribution function and $f \sim BART$. To estimate this model, we use the approach of Albert and Chib[22] and augment the model with latent variables Z_i:

$$\begin{aligned} Y_i &= I_{Z_i \geq 0} \\ Z_i &= \mu_0 + f(\boldsymbol{x}_i) + \epsilon_i \\ f(\boldsymbol{x}_i) &= \sum_{j=1}^{m} h(\boldsymbol{x}_i; T_j, M_j) \\ f &\sim BART \end{aligned} \tag{4}$$

where $I_{Z \geq 0}$ is one if $Z \geq 0$ and zero otherwise and $\epsilon_i \sim N(0,1)$. The Albert and Chib method then gives inference for f using the Gibbs sampler which draws $Z|f$ and $f|Z$.

The model just described can be readily estimated using existing software for binary BART. It allows one to estimate the functions $f(\boldsymbol{x})$ through Markov Chain Monte Carlo (MCMC) draws of f from which draws of the corresponding success probabilities $p(\boldsymbol{x}) = \Phi(\mu_0 + f(\boldsymbol{x}))$ are readily obtained. Here μ_0 is a tuning parameter to be chosen. For example, setting $\mu_0 = 0$ centers the prior for $p(\boldsymbol{x})$ at .5 since the BART prior for f is centered at 0; however, the BART model is fairly robust to different prior values of μ_0 given sufficient data. In the binary probit case, we let $\tau = 3/k$, where k is a tuning parameter. With this choice of τ and the default value of $k = 2$ recommended in literature[18] and used in the BART R package, there is a 95% prior probability that $f(\boldsymbol{x})$ is in the interval $\pm 2(1.5)$ giving a reasonable range of values for $p(\boldsymbol{x})$. Note that logit transformations could also be used for the binary outcome setting, and a logit implementation is also available in the BART package. However, because we are doing a prediction model and not focusing on parameter estimates like odds ratios, it is unclear whether probit or logit is more useful, so we have proceeded with the simpler probit framework.

4 BART for ITRs and value function estimation

Due to the excellent flexibility in modeling complex interactions and the strong predictive performance, we propose to use individualized predictions of patient outcomes from the BART model to determine an ITR. In addition to constructing an ITR based on BART predictions, we use the BART MCMC to assess uncertainty about various aspects of the treatment decision and resulting value function.

In applying BART to the ITR problem, our "$\boldsymbol{x}$" of the previous section (in which we reviewed BART) becomes (x, a) where, as in the section introducing ITR's, a is the decision variable and x is patient information.

The BART model implies $p(Y = 1|x, a, f) = \Phi(\mu_0 + f(x, a))$ where f is expressed as the sum of trees as in equation (4). Posterior MCMC draws $\{T_j^d, M_j^d\}$ consist of the individual trees $j = 1, 2, \ldots, m$, for MCMC iterations $d = 1, 2, \ldots, D$. Using equation (4), each of these draws results in a draw f_d of the function f. In this section, it is helpful to view the function f as the fundamental underlying parameter. BART is viewed as giving us draws $\{f_d\}_{d=1}^{D}$ from the posterior distribution of f.

4.1 Decision theoretic predictive BART ITR

From equation (1), the optimal ITR is obtained by choosing the value of a which maximizes $E(Y|x,a) = p(Y=1|x,a)$ conditional on $X=x$. In our Bayesian framework, $p(Y=1|x,a)$ is the predictive probability that $Y=1$ which is obtained by integrating $p(Y=1|x,a,f)$ over the posterior distribution of the parameter f. Given MCMC draws $\{f_d\}$ from the posterior distribution of f, this integral is approximated by averaging over the draws

$$p(Y=1|x,a) \approx \frac{1}{D}\sum_{d=1}^{D} p(Y=1|x,a,f_d) \equiv \bar{p}(x,a) \tag{5}$$

Thus, $\bar{p}(x,a)$ is the MCMC estimate of the predictive $p(Y=1|x,a)$. Illustrations of inference on patient specific predictive probabilities are shown later in section 8 and in Figure 5.

We can now define the BART-based ITR as the one in which the treatment for each individual is given by maximizing the patient specific predictive probability over available treatments

$$g_{\text{BART}}(x) = \arg\max_a \bar{p}(x,a) \tag{6}$$

The construction of g_{BART} follows the basic prescription of Bayesian decision theory in which we pick the action which maximizes expected utility. In this case, our utility is the outcome Y, and because Y is binary, the expected Y is the probability that $Y=1$, so we simply pick the action which gives us the highest predictive probability of a successful outcome.

4.2 Posterior distribution of the value function of an ITR

For any ITR $g(x)$, it is of interest to assess its value across the patient population so that different ITRs may be compared via this value. With an underlying function f, the value of the ITR g is defined as

$$V(g,f) = E_X(p(Y=1|x,g(x),f)$$

which is the average (over x) of the probability of a good outcome. Given MCMC draws, we can approximate the marginal distribution of this function of the uncertain f by simply plugging in draws of f

$$V_d(g) = V(g,f_d) \tag{7}$$

The posterior samples $\{V_d(g)\}$, $d=1,2,\ldots,D$, provide inference for the value function of any ITR g, including the BART ITR in equation (6), approximating the expectation over X by an average over a representative distribution of x which is often taken to be the observed samples in the covariate space. Illustrations of inference on the value function can be found later in section 8 and in Figure 5.

4.3 Posterior distribution of the value function of the optimal ITR

It is also possible to estimate and assess uncertainty of the value associated with the optimal ITR as defined in equation (1). We consider the optimal ITR as a function of f. If we knew f then, given x, the optimal action is given by

$$a(x,f) = \arg\max_a p(Y=1|x,a,f)$$

with corresponding maximum success probability $p^*(x,f) = \max_a p(Y=1|x,a,f)$. Draws of the value of the optimal ITR, namely $V^* = E_X(p^*(x,f))$ can be obtained from draws of f as

$$V_d^* = E_X(p^*(x,f_d)), \quad d=1,\ldots,D$$

Again, the expectation over X is typically an average of representative x values, often using the observed samples in the covariate space. We can interpret the draws V_d collectively as representing uncertainty about the value of the optimal ITR, i.e., the ITR for an agent who acts knowing the true f. Illustration of this is also found later in section 8 and Figure 5(c) and (d).

5 Comparison with existing methods

We conducted extensive simulation studies benchmarking the proposed BART ITR strategy against existing methods for identifying ITRs. In order to avoid any concerns that we are deliberately selecting simulation settings where BART will show good performance, we reproduced simulation settings from two recent papers[4,14] with a binary outcome variable Y and binary treatment A. The first set of simulations from a study[14] included five additional binary covariates $X_A : X_E$, five ordinal covariates $X_a : X_e$ with four categories each, and one or two continuous covariates X_{Ca}, X_{Cb}. Eight different scenarios for the logit of the probability of response were simulated according to the following

(A) $0.5X_{C1} + 2(X_{B1} + X_{a3} * X_{A1}) * A$
(B) $0.5X_{C1} + 2(X_{B1} + X_{a3} * (X_{b2} + X_{b3})) * A$
(C) $0.05(-X_{A1} + X_{B1}) + [(X_{a2} + X_{a3}) + (X_{b2} + X_{b3}) * X_{Ca}] * A$
(D) $\log\log[(X_{b3} + X_{c3}) + 5(X_{a2} + X_{a3} + X_{A1}X_{B1}) * A + 20]^2$
(E) $(X_{A1} + X_{B1}) + 2 * A$
(F) $0.5X_{A1} + 0.5X_{B1} + 2I(X_{Ca} < 5, X_a < 2) * A$
(G) $0.5X_{A1} + 0.5X_{B1} + 2I(X_{Ca} < 5, X_{Cb} < 2) * A$
(H) $0.5X_{Ca} + 0.5X_{Cb} + 2I(X_{Ca} < -2, X_{Cb} > 2) * A$

We also considered a modification of these scenarios (denoted (A2-H2)) in which the treatment interaction term was reduced to 1/4th of the given value, in order to better differentiate among the competing methods. In the second set of simulations from Kang et al.,[4] up to three independent continuous markers X_1, X_2, X_3 were included in seven different models for the probability of nonresponse according to the following

(K1) $\text{logit}p(Y = 0|A, X) = 0.3 + 0.2X_1 - 0.2X_2 - 0.2X_3 + A(-0.1 - 2X_1 - 0.7X_2 - 0.1X_3)$, $X_j \sim N(0, 1)$
(K2) $\text{logit}p(Y = 0|A, X) = 0.3 + 0.2X_1 - 0.2X_2 - 0.2X_3 + A(-0.1 - 2X_1 - 0.7X_2 - 0.1X_3)$, $X_j \sim N(0, 1)$ except for 2% of high leverage points with $X_1 \sim Uniform(8, 9)$.
(K3) $\log(-\log p(Y = 0|A, X)) = -0.7 - 0.2X_1 - 0.2X_2 + 0.1X_3 + A(0.1 + 2X_1 - X_2 - 0.3X_3)$, $X_j \sim N(0, 1)$
(K4) $\log(-\log p(Y = 0|A, X)) = 2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + A(-0.1 - X_1 + X_2)$, $X_j \sim Uniform(-1.5, 1.5)$
(K5) $\text{logit}p(Y = 0|A, X) = -0.1 - 0.2X_1 + 0.2X_2 - 0.1X_3 + X_1^2 + A(-0.5 - 2X_1 - X_2 - 0.1X_3 + 2X_1^2)$, $X_j \sim N(0, 1)$
(K6) $\text{logit}p(Y = 0|A, X) = 0.1 - 0.2X_1 + 0.2X_2 - X_1X_2 + A(-0.5 - X_1 + X_2 + 3X_1X_2)$, $X_j \sim N(0, 1)$
(K7) $p(Y = 0|A, X) = I(X_1 < 8)(1 + e^{-\eta})^{-1} + I(X_1 \geq 8)(1 - (1 + e^{-\eta})^{-1})$, where $\eta = 0.3 + 0.2X_1 - 0.2X_2 - 0.2X_3 + A(-0.1 - 2X_1 - 0.7X_2 - 0.1X_3)$ and $X_j \sim N(0, 1)$ except for 2% of high leverage points with $X_1 \sim Uniform(8, 9)$.

In all cases, ITRs were generated using a training dataset with $n = 500$ observations, and then each ITR was applied to a fixed independent test dataset of 2000 observations in order to compute the value function for this ITR from the true model. This process was replicated using 50 training datasets, and the average value function across the 50 training sets was obtained. This average value function for a particular ITR was normalized as a fraction of the true optimal value function to facilitate comparisons across scenarios.

For the BART ITR, we considered both use of the default prior parameters (BARTd), as well as cross-validation to select the number of trees (m = 80,200) and the value of k from among values of (0.2, 0.8, 2.0) with default value of 2 (BARTcv). Several competing methods were included for comparison, including regularized outcome weighted subgroup identification (ROWSI),[14] Outcome Weighted Learning (OWL),[16] use of random forests (RF) for outcome prediction along the lines of virtual twins approach[5] with cross validation of number of trees and minimum node size, and boosting with classification tree working model (KANG).[4] Ordinal variables were handled as ordinal for BART and other tree-based methods, but were otherwise treated as categorical variables.

Results are shown in Figure 2. In all cases, the value function of the BART ITR with cross-validation performed at or near the top of the competing methods. BART with the default settings also performed comparably to the other existing methods, with good performance in most situations.

6 Illustration of operating characteristics of BART prediction models and estimation of optimal value function

We demonstrate the features of the proposed method using generated data from two settings: a complex treatment interaction setting, and a main effect only setting. In each case, training datasets of either $n = 500$ or $n = 5000$ were

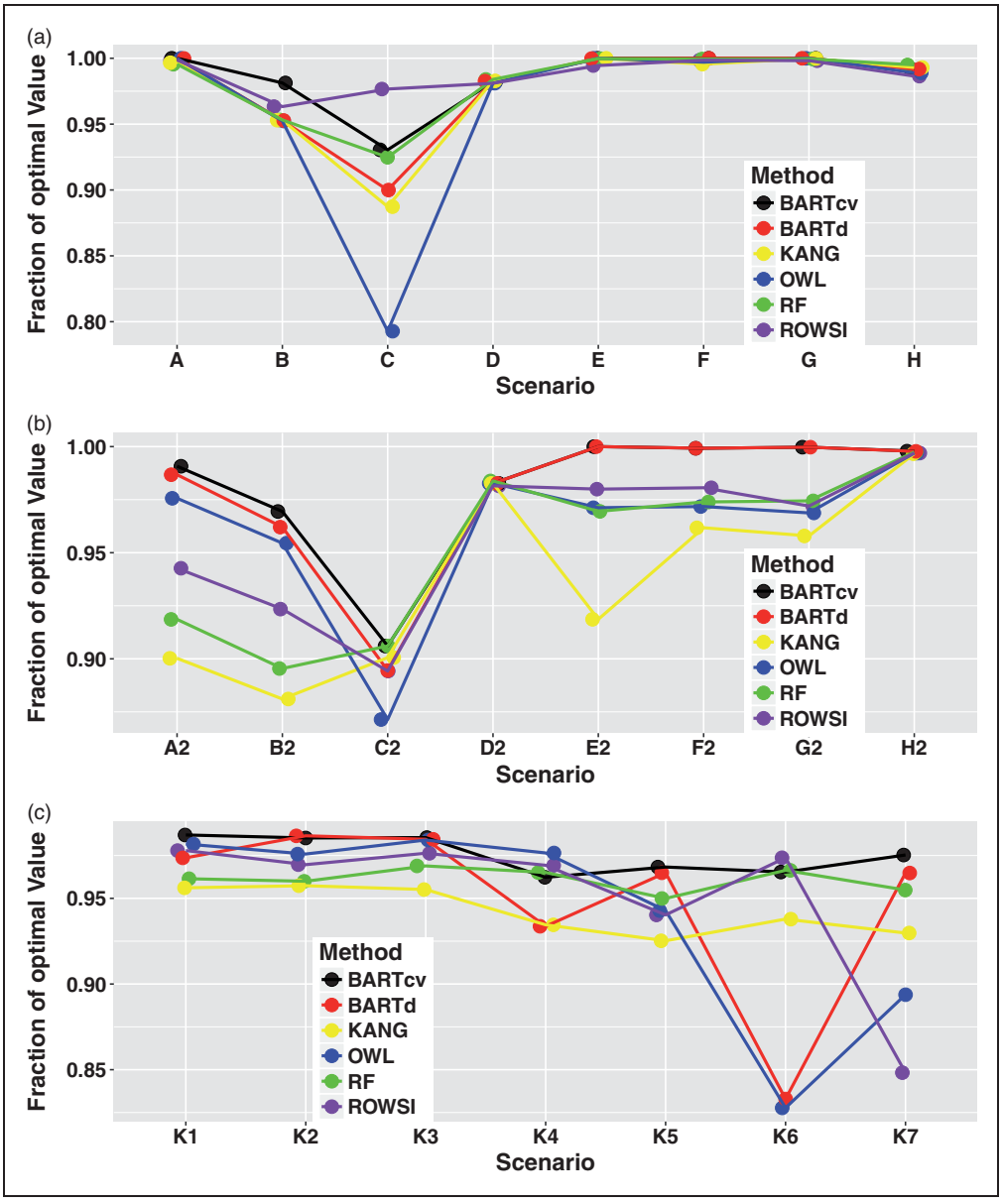

Figure 2. Value function relative to the optimal value function for simulation settings in (a) scenarios as in Xu et al.,[14] (b) same as (a) but with treatment effects cut by 25%, and (c) scenarios as in Kang et al.[4]

generated and applied to an independent test dataset of 2000 observations. Logistic regression with a binary outcome and three independent Uniform $(-1.5, 1.5)$ covariates was used to generate the data, with a complex treatment interaction model according to

$$P(Y=1|A,X) = \big[1+\exp\big\{-0.1-0.2X_1+0.2X_2-0.1X_3+0.5X_1^2 \\ +A(-0.5-0.5X_1-X_2-0.3I(X_3>0.5)+0.5X_1^2)\big\}\big]^{-1}$$

and a no treatment interaction model according to

$$P(Y=1|A,X) = \big[1+\exp\{-0.1-0.2X_1+0.2X_2-0.1X_3+0.5X_1^2-0.3A\}\big]^{-1}$$

In Figure 3, we plot the BART posterior means vs. the true probabilities of treatment outcome for the test dataset, for each treatment as well as for the treatment difference, for single training datasets of size $n=500$

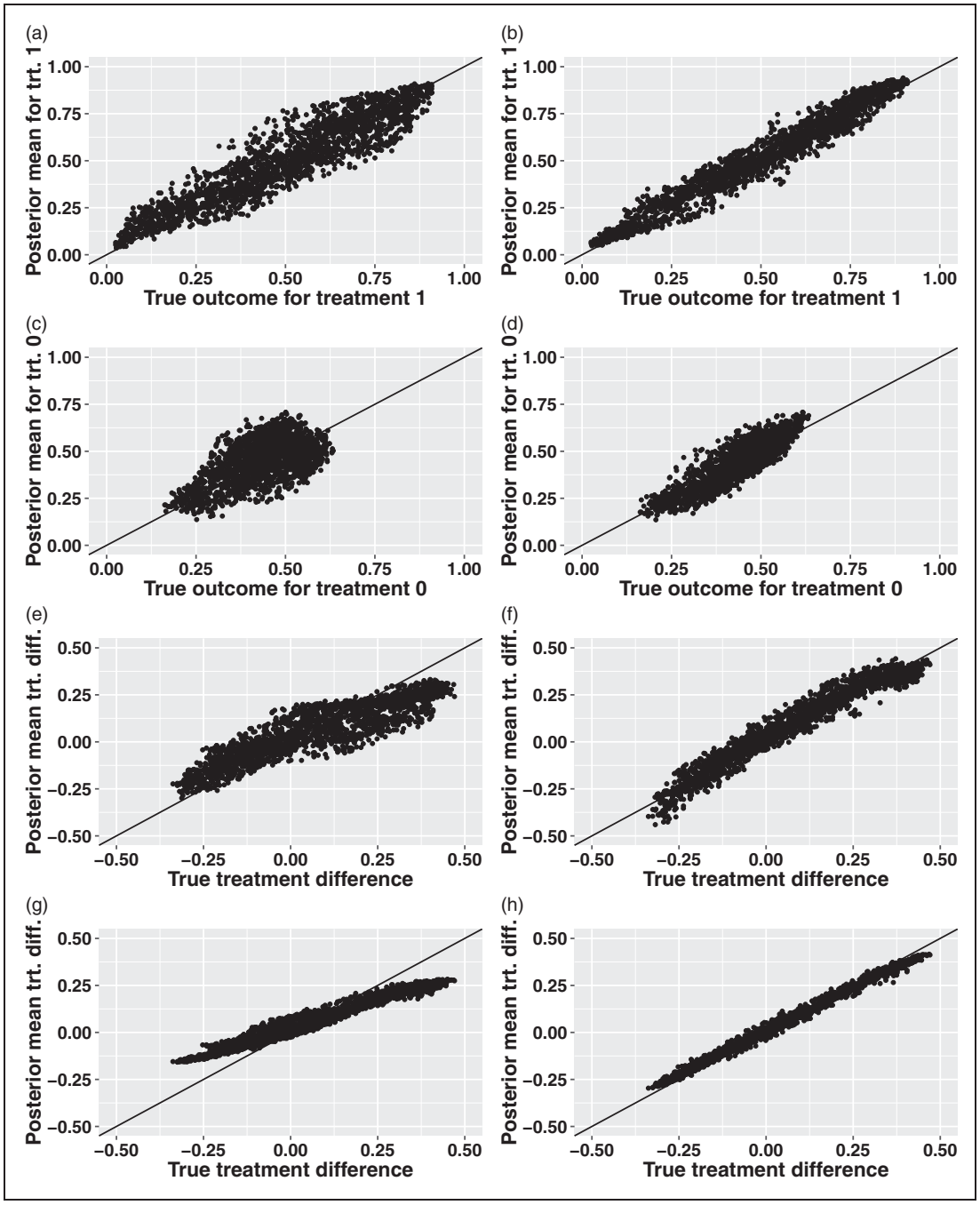

Figure 3. BART posterior means vs. true probabilities for complex interaction model, with $n = 500$ (a,c,e,g) and $n = 5000$ (b,d,f,h). (a,b,c,d) show predictions for individual treatment outcomes, while (e,f) show predictions for treatment differences, all using a single training dataset. (g,h) show the posterior means for the treatment difference averaged over 400 repeated data simulations of the training set.

((a),(c),(e)) and $n = 5000$ ((b),(d),(f)) using the complex interaction model. The posterior mean prediction from the BART model has excellent accuracy for the large sample size, both for individual treatment outcomes as well as for the treatment difference, despite the complex treatment interaction model which includes linear and quadratic covariate-treatment interactions as well as an interaction term with a thresholded value of a covariate. The larger training dataset shows improved accuracy compared to the smaller training dataset, which has some modest shrinkage of the treatment effects.

We also conducted repeated data simulations with 400 replicates to look at the bias of the prediction model as well as coverage of the 95% interval estimates for the value function of the optimal ITR, using the quantiles of the posterior samples for the value function of the optimal ITR. Bias over 400 replicates for the treatment difference is shown in Figure 3(g) and (h). As is seen with the single training dataset, there is some small bias and shrinkage of the treatment effects for smaller sample size which disappears with larger sample size. Coverage probabilities of 95% credible intervals for the value of the optimal ITR are 90% for training dataset of size $n = 500$ and 95% for training dataset of size $n = 5000$, indicating that once the sample size is sufficient to reduce the shrinkage of the treatment effects, coverage of the value function for the optimal ITR is excellent. We also examined coverage of 95% credible intervals for the treatment effect for each individual in the test dataset; median (inter-quartile range) of these coverage probabilities are 96.7% (89.5%–99.0%) for training datasets of size $n = 500$, and 99.2% (98.5%–99.5%) for training datasets of size $n = 5000$. Average widths of these 95% credible intervals for the treatment effect for each individual were 0.43 for $n = 500$ and 0.25 for $n = 5000$.

Similar results are shown for the model with no treatment-covariate interaction in Figure 4. Note that the true treatment differences show very narrow variability due to the data generating model, and the BART model shows predictions for treatment differences which are also small and which narrow with increasing sample size. These predictions are unbiased in repeated simulation, as indicated by the convergence to the diagonal line in Figure 4(c) and (d). Coverage of the value function and individual treatment effects was similar to the complex interaction setting.

7 Summarizing the BART ITR

The ITR based on the BART prediction model does not directly yield a simple interpretable rule; this issue in general with flexible models has been discussed in literature,[7] who propose directly optimizing the value function over an interpretable set of rules. In contrast, we separate the modeling of outcome from the determination of an interpretable rule, by trying to develop an approximation to this BART ITR which is interpretable and yields good performance. We propose a ''fit-the-fit'' strategy, in which one develops a single tree fit to the posterior mean treatment differences as a function of patient characteristics. Essentially, the posterior mean treatment differences are treated as the ''data,'' and we try and fit an interpretable single tree to this data. The single tree then provides an interpretable way to explain which groups of patients should receive which treatment, as well as the magnitude of the treatment difference for that group of patients. This strategy was originally proposed as a variable selection technique for a BART prediction model, but has been adapted here to focus on summarizing the inference on treatment differences and the BART ITR. A similar strategy of applying a tree to estimated treatment differences was proposed by Foster et al.,[5,23] where they also note that estimation of individual treatment effects using random forests could be improved by expanding the set of predictor variables to explicitly include treatment interactions. Expansion of the set of predictor variables could also be used in BART, though any improvements would likely need inclusion of the correct form of the interactions which may be difficult in practice. Also, there may be a cost to adding a lot of additional terms to be considered in the trees. Further research is needed to investigate the potential benefit of this strategy.

To fit an appropriate tree and also identify the best set of variables to include in that tree, a sequence of trees are fit, where variables are sequentially added to the candidate set of splitting variables in a stepwise manner to improve the fit. Given a current set of variables and a corresponding tree built using those variables, the fit of the current tree is assessed using the R^2 between the fitted values from the tree and the posterior mean treatment difference (which is being used as the ''data''). Additional variables are considered to be added one at a time yielding new trees built with an increasing set of variables, and their R^2 is assessed for each new variable added and corresponding tree. The variable which most increases the R^2 is selected at each step. Once the R^2 does not improve appreciably with addition of a new variable (we used $\Delta R^2 < 1\%$), the procedure ends and the current tree is used as the approximation to the BART ITR.

Note that each single tree fit can be implemented very quickly, so this fit the fit postprocessing procedure takes minimal computing time. It may not always be possible to identify such a single tree (this is in fact the benefit of ensemble methods to provide improved prediction), but note that the quality of such an interpretable approximation can be assessed using the R^2 between the single tree and the BART prediction model. It is also possible to compute the value function of the single-tree approximation ITR, and compare it with that of the full BART-ITR.

This ''fit-the-fit'' approach may require a very complicated tree in order to provide a sufficiently accurate approximation to the BART treatment differences, and sometimes it still may be unsuccessful; this is precisely the situation where an ensemble prediction model is most needed. A complicated tree may also lose ease of

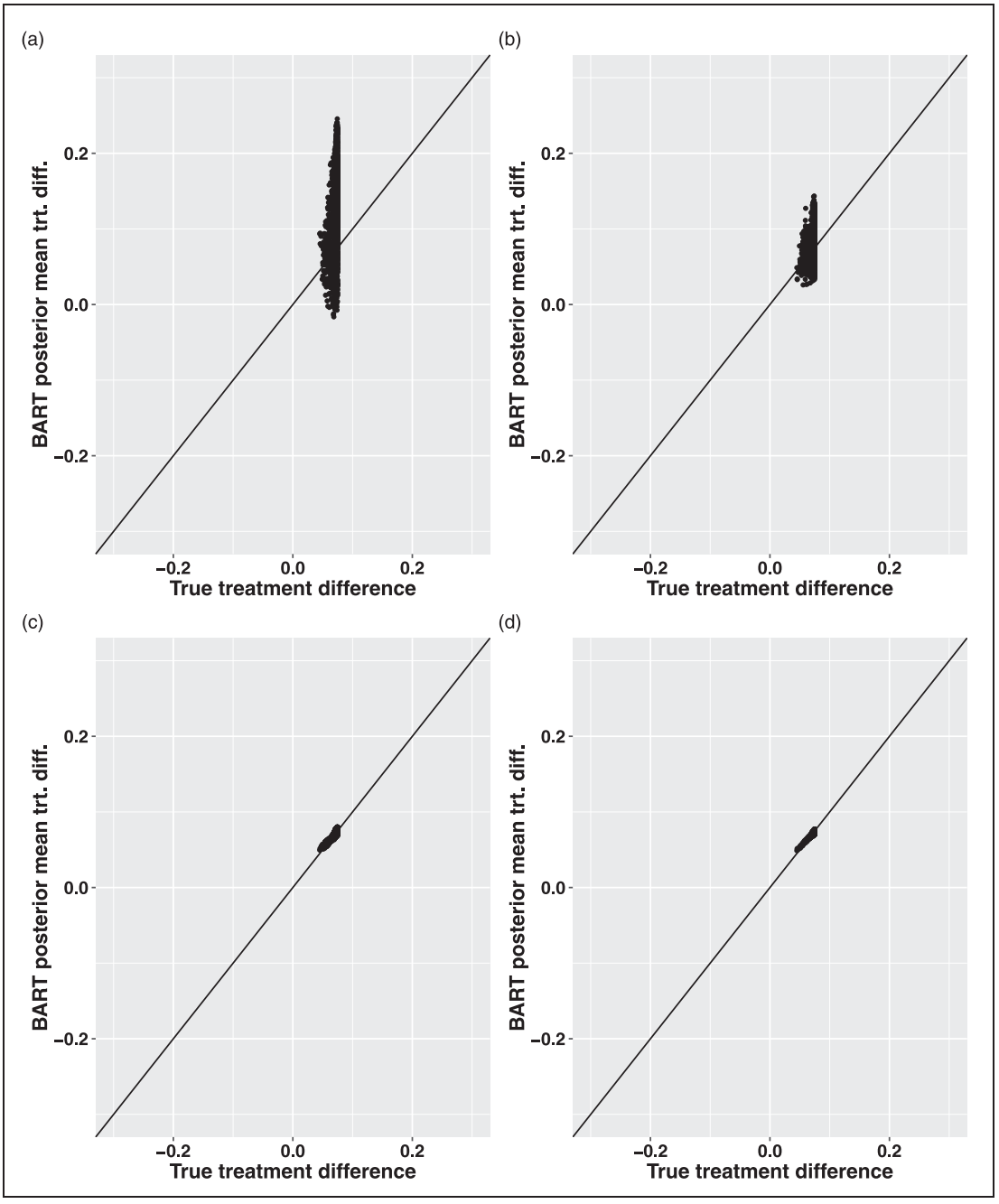

Figure 4. BART posterior means vs. true probabilities for no interaction model, with $n = 500$ (a,c) and $n = 5000$ (b,d). (a,b) show predictions for treatment differences using a single training dataset, while (c,d) show the posterior means for the treatment difference averaged over 400 repeated data simulations of the training set.

interpretability. As an alternative, one also could consider growing a classification tree to fit the treatment decision process, rather than the treatment effect itself. This may result in a simpler tree which is accurate for approximating BART-based treatment decisions, but does not necessarily differentiate between small and large treatment effects.

8 Example

We illustrate the proposed methodology with a study of one year survival outcomes after a hematopoietic cell transplant (HCT) used to treat a variety of hematologic malignancies. We use data of 3802 patients receiving

reduced intensity conditioning for their HCT between 2011 and 2013, with data reported to the Center for International Blood and Marrow Transplant Research (CIBMTR). Follow up to one year is complete for all patients, so we analyze the outcomes using binary methods as described throughout the paper applied to the patient's survival status at one year. Note that a full survival analysis could also be conducted with BART methods available for survival data,[24] however this requires consideration of the appropriate target outcome to optimize and so we defer this for future research. The primary treatment of interest is the type of conditioning regimen used (Fludarabine/Melphalan, or FluMel for short, vs. Fludarabine/Busulfan, or FluBu for short). A variety of patient, donor, and disease factors were examined for their utility in personalizing the selection of the conditioning regimen, including age, race/ethnicity, performance score, Cytomegalovirus status, disease, remission status, disease subtypes, chemosensitivity, interval from diagnosis to transplant, donor type, Human Leukocyte Antigen (HLA) matching between donor and recipient, prior autologous transplant, gender matching between donor and recipient, comorbidity score, and year of transplant. This observational cohort appears to be well balanced between the regimens across these factors, indicating reasonable equipoise by clinicians on which conditioning regimen is most appropriate for individual patients.

Fitting of the BART model provides samples from $p(Y = 1|x, a, f_d)$ for $d = 1, \ldots, D$. In Figure 5 we show waterfall plots of the differences in one year survival between Flu/Mel and Flu/Bu conditioning across patients in two ways. In Figure 5(a) we use the samples from the patient specific difference in one year survival, $p(Y = 1|x, a = \text{Flu/Mel}, f_d) - p(Y = 1|x, a = \text{Flu/Bu}, f_d)$, and plot the posterior mean of these differences in one year survival for each patient, sorted by the magnitude of the difference. This is equivalent to the difference in predictive probabilities under each treatment condition as described in equation (5). Inter-quartile ranges and 95% posterior intervals are also shown to indicate the variability of the differences. In Figure 5(b) we show the waterfall plot of the posterior probabilities that Flu/Mel has a higher one year survival than Flu/Bu. This is obtained by computing

$$\frac{1}{D}\sum_{d} I(P(Y = 1|x, a = \text{Flu/Mel}, f_d) > P(Y = 1|x, a = \text{Flu/Bu}, f_d))$$

These plots indicate some heterogeneity in treatment benefit, where approximately 3000 of the patients seem to benefit from Flu/Mel, albeit with varying degrees of magnitude and/or certainty surrounding that benefit, while the remaining patients seem to benefit from Flu/Bu conditioning.

We implemented BART using the default settings of $m = 200$ and $k = 2$. To examine sensitivity of the width of the posterior intervals for the patient specific treatment difference to the choice of tuning parameters, we examined these widths under standard values of m and k. Average widths across patients in the cohort with $m = 50$ were 0.13, 0.13, 0.12 for $k = 1$, 2, 3, respectively. Average widths with $m = 200$ were 0.16, 0.14, 0.12 for $k = 1$, 2, 3, respectively. This suggests only mild sensitivity to the tuning parameters, with higher values of k and lower values of m tending to produce slightly narrower intervals, likely due to greater borrowing of information from neighboring observations. Note however that narrower intervals due to larger k or smaller m may also have greater bias due to this same phenomenon, so narrower interval widths are not the best target for choosing m and k. Cross-validation focused on prediction error is useful for this purpose, and for this dataset, cross validation ended up also selecting the default values of $k = 2$ and $m = 200$.

Also in Figure 5(c), we show the value functions as described in equation (7) for the cohort of $n = 3802$ patients for three treatment rules: all patients receive Flu/Bu, all patients receive Flu/Mel, and patients receive treatment according to the BART-based ITR. The BART ITR value function distribution is shifted to the right, indicating improved one year survival outcomes over the overall cohort using this individualized strategy. The posterior mean of the value function distributions for each treatment strategy are: FluBu: 0.651, FluMel: 0.667, BART ITR: 0.677, optimal ITR: 0.682. Figure 5(d) shows the density functions of the difference in value function for the BART ITR compared to the other strategies, indicating high likelihood that the BART ITR is superior to the fixed treatment strategies. On the other hand, it is inferior to the optimal ITR as expected, but the differences are small.

As pointed out earlier, one drawback of the BART-based ITR is that it does not lead to a simple interpretable rule. Next we apply the fit-the-fit technique to approximate the BART-based ITR with an interpretable treatment rule which has nearly as good performance. In order to do this, the posterior mean treatment differences in one year survival are treated as the "data," and we try and fit an interpretable single tree to this data. A sequence of trees are fit, where variables are added sequentially to the set of potential splitting variables of the tree in a stepwise manner to improve the fit. Once the change in R^2 is less than 1% with addition of a new variable, the procedure ends and the current tree is used as the approximation to the BART ITR. The results of the final tree fit are shown

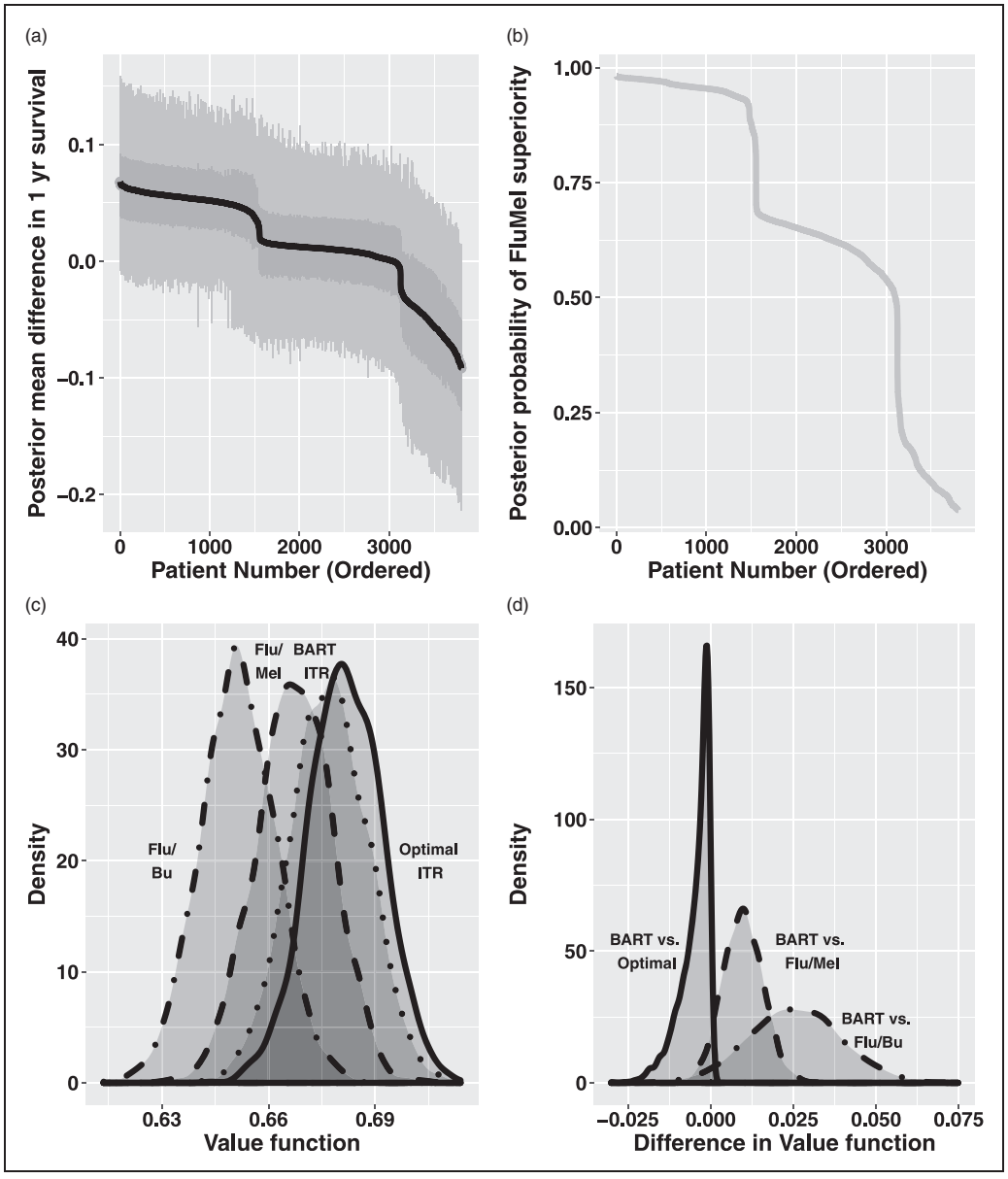

Figure 5. Results of BART ITR for HCT example. (a) Waterfall plot of one year survival differences (FluMel-FluBu) by patient (posterior mean differences, along with inter-quartile ranges and 95% posterior intervals), (b) Waterfall plot of posterior probabilities that survival is higher for Flu/Mel, (c) Density plot of value functions for three treatment strategies (FluMel, FluBu, BART ITR) as well as optimal ITR, and (d) Density plot of difference in value functions for treatment strategies compared to BART ITR. The posterior mean of the value function distributions for each treatment strategy are: FluBu: 0.651, FluMel: 0.667, BART ITR: 0.677, and optimal ITR: 0.682.

in Figure 6; R^2 between the tree fit and the posterior mean treatment differences is 97%. The first splitting variable used, Non-Hodgkin's Lymphoma (NHL) disease vs. other disease, was sufficient to match the BART ITR exactly in terms of selection of conditioning regimen; Patients with NHL have approximately 5% better one year OS with Flu/Bu conditioning, while patients without NHL have approximately 3% better one year OS with Flu/Mel conditioning. A second level of splitting variables provided further resolution on the magnitude of the treatment benefit for disease subgroups, but did not affect the directional benefit. The splits in the tree also match up with the drops in the treatment effects seen in the waterfall plot in Figure 5, where the drops from each plateau seem to indicate a subpopulation change. Patients with AML have approximately 5% better one year OS with Flu/Mel conditioning, while patients with other diseases have approximately 1% better one year OS with

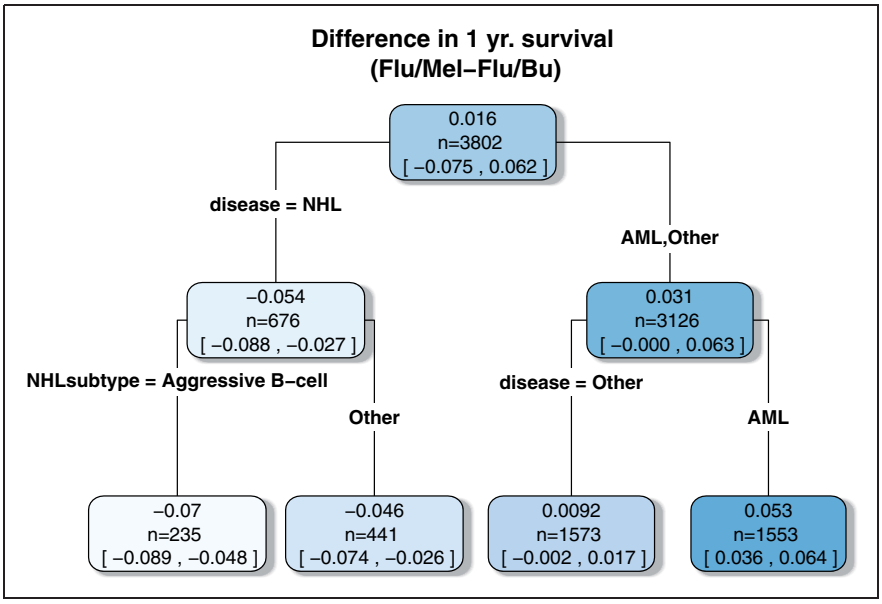

Figure 6. Tree fit to the posterior mean treatment differences. Values in each node represent the posterior mean and 95% credible intervals for the average treatment effect of the subgroup of patients represented in that node.

Flu/Mel conditioning. Similarly, among patients with NHL disease, those with aggressive B-cell subtype may experience slightly more benefit from Flu/Bu conditioning compared to those with other subtypes (7% vs. 5%). Note that in this case, this simple rule completely matched the BART-based ITR, and so the value function associated with this rule also matches the BART ITR value function in the previous figure. This type of approximation to a BART-based ITR can greatly facilitate communication with clinicians on how the ITR works and what types of patients benefit from one treatment vs. another.

9 Discussion

In this article, we presented a framework for identifying optimal individualized treatment rules using Bayesian Additive Regression Trees. BART is a flexible fully nonparametric prediction model which can handle complex functional forms as well as interactions among variables, and therefore it is well suited for examining treatment interactions which drive ITRs. There are two main advantages to using BART to identify an ITR. First, our proposed method has excellent performance when benchmarked against existing methods, including other flexible prediction methods as well as policy search methods such as outcome weighted learning; overall it performed better than or comparable to other existing methods across a range of ITR simulation settings established in other papers. Second, our method provides direct inference on the value of the BART ITR as well as the optimal ITR. This requires incorporation of both the uncertainty in the prediction model, as well as uncertainty in the individual patient treatment selection that depends on the prediction model. Both can be handled in a straightforward manner using the posterior samples for the prediction model function in the Bayesian framework. In contrast, it is not clear how to do this for policy search methods which do not provide a direct prediction model for outcome, as well as for other flexible models such as random forests which do not directly provide uncertainty measures.

We observed that the BART model tends to shrink the treatment effect estimates for smaller sample size leading to underestimation of the true value of the optimal ITR. Larger sample sizes do overcome this shrinkage from the prior. Further consideration of ways to reduce this unidirectional bias could provide better inference on the value function for small sample sizes.

One limitation of our proposed method is that the BART method generates a ''black box'' prediction model which is difficult to explain to clinicians. However, we showed how the posterior mean differences available from the model can be fed into a tree procedure to yield an interpretable ITR which provides a close approximation to the BART ITR and which can be readily explained to clinicians. A recent article[25] describes the use of BART and

utility specifications with the goal of identifying subgroups with elevated treatment effects, in contrast to finding an individualized treatment rule.

Our proposed method uses ''off-the-shelf'' BART software; minimal processing is needed to obtain posterior inference under each treatment condition for patients in a test dataset. Inference for the value of an ITR is also readily available. While BART can be computationally demanding as an MCMC technique, it can be parallelized to save computational time since the chains do not share information beyond the data itself. An example of code to implement the methods discussed in this manuscript is available in the Supplemental Material.

Acknowledgements

The authors would like to thank Dr Menggang Yu from University of Wisconsin for providing software to implement the ROWSI method in the simulation study.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported in part by the Advancing a Healthier Wisconsin Endowment at the Medical College of Wisconsin.

Supplemental material

Supplemental material is available online for this article.

References

1. Qian M and Murphy SA. Performance guarantees for individualized treatment rules. *Ann Stat* 2011; **39**: 1180–1210.
2. Imai K and Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* 2013; **7**: 443–470.
3. Moodie EE, Dean N and Sun YR. Q-learning: flexible learning about useful utilities. *Stat Biosci* 2014; **6**: 223–243.
4. Kang C, Janes H and Huang Y. Combining biomarkers to optimize patient treatment recommendations. *Biometrics* 2014; **70**: 695–707.
5. Foster JC, Taylor JMG and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; **30**: 2867–2880.
6. Zhao Y, Zeng D, Socinski MA, et al. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* 2011; **67**: 1422–1433.
7. Zhang Y, Laber EB, Tsiatis A, et al. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* 2015; **71**: 895–904.
8. Dusseldorp E, Conversano C and Os BJV. Combining an additive and tree-based regression model simultaneously: STIMA. *J Comput Graph Stat* 2010; **19**: 514–530.
9. Doove LL, Dusseldorp E, Deun KV, et al. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classif* 2014; **8**: 403–425.
10. Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 2011; **30**: 2601–2621.
11. Su X, Tsai CL, Wang H, et al. Subgroup analysis via recursive partitioning. *J Mach Learn Res* 2009; **10**: 141–158.
12. Zeileis A, Hothorn T and Hornik K. Model-based recursive partitioning. *J Comput Graph Stat* 2008; **17**: 492–514.
13. Laber E and Zhao Y. Tree-based methods for individualized treatment regimes. *Biometrika* 2015; **102**: 501–514.
14. Xu Y, Yu M, Zhao YQ, et al. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* 2015; **71**: 645–653.
15. Zhang B, Tsiatis AA, Laber EB, et al. A robust method for estimating optimal treatment regimes. *Biometrics* 2012; **68**: 1010–1018.
16. Zhao Y, Zeng D, Rush AJ, et al. Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc* 2012; **107**: 1106–1118.
17. Fu H, Zhou J and Faries DE. Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Stat Med* 2016; **35**: 3285–3302.

18. Chipman HA, George EI and McCulloch RE. Bart: Bayesian Additive Regression Trees. *Ann Appl Stat* 2010; **4**: 266–298.
19. Linero A. Bayesian regression trees for high dimensional prediction and variable selection. *J Am Stat Assoc* 2017. http://dx.doi.org/10.1080/01621459.2016.1264957 (accessed 24 November 2017).
20. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 2011; **20**: 217–240.
21. Gelfand AE and Smith AF. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 1990; **85**: 398–409.
22. Albert J and Chib S. Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* 1993; **88**: 669–679.
23. Foster JC, Taylor JM, Kaciroti N, et al. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics* 2015; **16**: 368–382.
24. Sparapani RA, Logan BR, McCulloch RE, et al. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Stat Med* 2016; **35**: 2741–2753.
25. Sivaganesan S, Müller P and Huang B. Subgroup finding via Bayesian Additive Regression Trees. *Stat Med* 2017. http://dx.doi.org/10.1002/sim.7276 (accessed 24 November 2017).