Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Estimation of causal effects of multiple treatments in observational studies with a binary outcome

Statistical Methods in Medical Research
2020, Vol. 29(11) 3218–3234

DOI: 10.1177/0962280220921909
journals.sagepub.com/home/smm
SAGE

Liangyuan Hu[1,2,3], Chenyang Gu[4], Michael Lopez[5], Jiayi Ji[1,2,3] and Juan Wisnivesky[6]

Abstract
There is a dearth of robust methods to estimate the causal effects of multiple treatments when the outcome is binary. This paper uses two unique sets of simulations to propose and evaluate the use of Bayesian additive regression trees in such settings. First, we compare Bayesian additive regression trees to several approaches that have been proposed for continuous outcomes, including inverse probability of treatment weighting, targeted maximum likelihood estimator, vector matching, and regression adjustment. Results suggest that under conditions of non-linearity and non-additivity of both the treatment assignment and outcome generating mechanisms, Bayesian additive regression trees, targeted maximum likelihood estimator, and inverse probability of treatment weighting using generalized boosted models provide better bias reduction and smaller root mean squared error. Bayesian additive regression trees and targeted maximum likelihood estimator provide more consistent 95% confidence interval coverage and better large-sample convergence property. Second, we supply Bayesian additive regression trees with a strategy to identify a common support region for retaining inferential units and for avoiding extrapolating over areas of the covariate space where common support does not exist. Bayesian additive regression trees retain more inferential units than the generalized propensity score-based strategy, and shows lower bias, compared to targeted maximum likelihood estimator or generalized boosted model, in a variety of scenarios differing by the degree of covariate overlap. A case study examining the effects of three surgical approaches for non-small cell lung cancer demonstrates the methods.

Keywords
Causal inference, generalized propensity score, inverse probability of treatment weighting, matching, machine learning

1 Introduction

1.1 Motivating research question

Lung cancer is the leading cause of cancer-related mortality worldwide and is estimated to have caused over 1.7 million deaths in 2018.[1] The most common type of lung cancer is non-small cell lung cancer (NSCLC), accounting

[1]Department of Population Health Science and Policy, Icahn School of Medicine, New York, USA
[2]Institute for Health Care Delivery Science, Icahn School of Medicine, New York, USA
[3]Tisch Cancer Institute, Icahn School of Medicine, New York, USA
[4]Analysis Group, Inc., Los Angeles, USA
[5]National Football League, New York, USA
[6]Department of Medicine, Icahn School of Medicine, New York, USA

Corresponding author:
Liangyuan Hu, Center for Biostatistics, Department of Population Health Science and Policy, Ichan School of Medicine, New York, NY 10029, USA.
Email:liangyuan.hu@mssm.edu

for approximately 85% of all lung cancer cases.[2] When feasible, NSCLC tumors are treated using surgical resection, which remains the most effective option for a cure.[3]

Open thoracotomy long stood as the standard surgical procedure for stage I–IIIA NSCLC tumors. However, open thoracotomy is associated with considerable postoperative complications and mortality, especially in the elderly.[4,5] Beginning in the late 1990s, two newer and less invasive techniques, video-assisted thoracic surgery (VATS) and, more recently, robotic-assisted surgery, were increasingly used.[6,7] The adoption of VATS and robotic-assisted surgery seemed to signal that the newer procedures offer a clinical benefit relative to open resection.[8,9] However, to our knowledge, no randomized controlled trials (RCTs) have been conducted to compare the effectiveness of these surgical procedures, in part due to difficulties in recruiting patients and high study costs. As a consequence, VATS and robotic-approaches were adopted into routine care without sufficient scrutiny.[6,10]

In place of RCTs, large-scale population-based databases, such as the Surveillance, Epidemiology, and End Results (SEER)-Medicare database, provide research opportunities for comparative studies. The SEER-Medicare database comprises a large sample of patients who received each of the three surgical procedures and reflects patient outcomes in the real world setting, containing demographic and clinical information for Medicare beneficiaries with cancer in various United States regions.[11] However, in contrast to RCTs, the real-world adoption pattern of the three surgical approaches largely depends on the patients' sociodemographic and tumor characteristics, which may result in an unbalanced cohort with significant differences in the distributions of sociodemographic characteristics, comorbidities, cancer characteristics, and diagnostic information across treatment groups.[12]

The research question poses several challenges for statistical analyses. First, in practice, statistical methods designed for a binary treatment are often used to account for underlying differences in patient characteristics to compare each pair of surgical procedures.[6,10,13] Unfortunately, applications of these methods can lead to the comparisons of disparate patient subgroups, which may increase bias in treatment effect estimates.[14] Second, common measures for comparative effectiveness are postoperative complications, which are binary outcomes. Thus, the treatment effects are typically based on the risk difference (RD), odds ratio (OD), or relative risk (RR),[15] all of which make it less straightforward to obtain inference, relative to continuous outcomes.[16–18] Third, the robotic-assisted surgery is a new advanced technology that was just adopted into practice in recent years. As a result, the number of patients who are operated via this approach is smaller compared to the other two approaches, yielding unequal sample sizes across the treatment groups. Appropriate causal inference methods that can address these challenges are needed.

1.2 Overview of methods for causal inference with multiple treatments

Recent years have seen a growing interest in the development of causal inference methods with multiple treatments using observational data. The theoretical work of Imbens[19] and Imai and Van Dyk[20] extended the propensity score framework in the setting with a binary treatment[21] to the general treatment setting. Subsequently, methods designed for a binary treatment have been reformulated to accommodate multiple treatments, including regression adjustment (RA),[22] inverse probability of treatment weighting (IPTW),[23,24] and vector matching (VM).[14] Lopez and Gutman[14] provide a comprehensive review of current methods for multiple treatments. These methods focus on continuous outcomes.

RA,[22,25,26] also known as model-based imputation,[27] uses a regression model to impute missing outcomes, estimating what would have happened to a specific unit had this unit received the treatment to which it was not exposed. The causal estimand of interest can be estimated by contrasting the imputed potential outcomes between treatment groups. The critical part of this method is the specification of the functional form of the regression model. With a low-dimensional set of pre-treatment covariates, it is relatively easy to specify a flexible functional form for the regression model. If there are many pre-treatment covariates, however, such a specification is more difficult, and possible misspecification of the regression model could bias the estimate of treatment effects. RA also heavily relies on extrapolation for estimation when the covariate distributions between treatment groups are far apart.[27]

IPTW[19,23,24] methods attempt to obtain an unbiased estimator for treatment effects in a way akin to how weighting by the inverse of the selection probability adjusts for unbalances in sampling pools, introduced by Horvitz and Thompson[28] in survey research. A challenge with IPTW is that treated units with low generalized propensity scores (GPSs) that are close to zero can result in extreme weights, which may yield erratic causal estimates with large sample variances.[29,30] This issue is increasingly likely as the number of treatments increases.[14]

An alternative method is to use trimmed or truncated weights, in which weights that exceed a specified threshold are each set to that threshold.[31,32] The threshold is often based on quantiles of the distribution of the weights (e.g. the 1st and 99th percentiles).

Alternatives to estimate GPSs in the IPTW framework include generalized boosted models[24] (GBMs) and Super Learner[33,34] (SL). GBMs grow multiple regression trees to capture complex and nonlinear relationships between treatment assignment and pre-treatment variables. The estimation procedure can be tuned to find the GPS model producing the best covariate balance between treatment groups. This feature of GBMs should help alleviate extreme weights and improve the estimation of causal effects.[24] However, the algorithm can be computationally intensive, and the robust procedure for estimating the variances of the effect estimates is not guaranteed to result in proper confidence intervals. SL uses ensemble of machine learning approaches including regression, ridge regression, and classification trees, to estimate a weight for each treatment. There is no guarantee that these probabilities sum to 1, and it is common to normalize weights accordingly. To limit extreme weights, Rose and Normand[34] use a lower bound of 0.025 for each probability.

Targeted maximum likelihood estimation[34,35] (TMLE) is a doubly robust approach that combines outcome estimation, IPTW estimation, and a targeting step to optimize the parameter of interest with respect to bias/variance. Rose and Normand[34] implement TMLE by estimating both GPSs and a binary outcome using SL. For obtaining variance estimates, Rose and Normand[34] use influence curves, though bootstrapping is also suggested. The use of TMLE has, to the best of our knowledge, not been deeply vetted for multiple treatment options by using simulations.

Lopez and Gutman[14] proposed the VM algorithm, which can match units with similar vector of GPSs. VM is designed to replicate a multi-arm randomized trial by generating sets of units that are roughly equivalent on measured pre-treatment covariates. VM obtains matched sets using a combination of k-means clustering and one-to-one matching with replacement within each cluster strata. Simulations demonstrated that, relative to IPTW with the GPSs estimated using multinomial logistic regression, and to generalizations of tools designed for binary treatment, VM yielded lower bias in the covariates' distributions between different treatment groups, while retaining most of eligible units that received the reference treatment.[14] However, the authors' acknowledge that there is a lack of guidance regarding the estimation of the sampling variance, and this is an area for further statistical research.

Before describing Bayesian additive regression trees (BART)—one tool that we think is equipped to handle the complexity of causal inference with multiple treatments—it is worth explicating on why one approach that, although intuitive and easy to perform, is not recommended: a series of binary comparisons (SBCs). To wit, grouping subjects into separate sub-populations, each with two treatments, and then using approaches designed for binary treatment is an approach often used in practice.[14] However, SBCs can (i) lead to non-transitive causal estimates, (ii) increase bias, and (iii) leave it unclear which treatment is optimal, all of which make it inappropriate for causal inference when there are more than two treatments.[14]

1.3 BART for causal inference

While the advanced regression and propensity score-based techniques described above were created for causal inference with multiple treatments, these methods were developed with continuous outcomes in mind, and they have been less studied in the context of both a binary outcome and multiple treatments.

In recent years, BART,[36,37] a nonparametric modeling tool, has become more popular in causal settings. Hill[38] proposed the use of BART for causal inference with a binary treatment and a continuous outcome. Hill[38] and Hill and Su[39] used simulations to show that, in scenarios where there are non-linearities in the response surface and the treatment assignment mechanism, BART generates more accurate estimates of average treatment effects (ATEs) compared to various matching and weighting techniques, and comparable estimates in linear settings.

BART boasts several advantages for causal inference with a binary treatment.[38,40] First, BART allows for an extremely flexible functional form. Second, BART avoids ambiguity with respect to covariate balance diagnostics required by propensity score-based approaches. Third, BART generates coherent uncertainty intervals for treatment effect estimates from the posterior samples in contrast to propensity score matching and subclassification, for which there is lack of agreement regarding appropriate interval estimation.[20,38] Finally, BART is easy to implement and requires less researcher programming expertise. However, like any methods that do not first discard units that fall out of areas of the covariate space where common support does not exist, one vulnerability of BART is that there is no mechanism to prevent it from extrapolating over these areas.

We surmise that the strengths of BART are transferable to the multiple treatment setting. In the sections that follow, we conduct two sets of simulations to investigate the operating characteristics of BART for estimating the causal effects of multiple treatments on a binary outcome, and compare BART to the existing methods discussed previously. We further supply BART with a strategy to identify a common support region and compare it to the propensity score-based strategy with respect to the proportion of units retained for inference and the accuracy of treatment effect estimates based on the retained inferential units. We subsequently apply the methods examined to analyze a large dataset on stage I–IIIA NSCLC patients, drawn from the SEER-Medicare registry, and estimate the comparative effect of robotic-assisted surgery versus VATS and open thoracotomy on postoperative outcomes.

2 Potential outcomes framework for multiple treatments

2.1 Notation and assumptions

Our notation is based on the potential outcomes framework, which was originally proposed by Neyman[41] in the context of randomization-based inference in experiments. Potential outcomes were generalized to observational studies and Bayesian analysis by Rubin,[42–44] in what is now known as the Rubin causal model.[45]

Consider a sample of N units, indexed by $i = 1, \dots, N$, drawn randomly from a target population, which comprises individuals in a study designed to evaluate the effect of a treatment W on some outcome Y. Each unit is exposed to one of total Z possible treatments, that is $W_i = w$ if individual i was observed under treatment w, where $w \in \mathcal{W} = \{1, 2, \dots, Z\}$. The number of units receiving treatment w is n_w, where $\sum_{w=1}^{Z} n_w = N$. For each unit i, there is a vector of pre-treatment covariates, $\boldsymbol{X}_i$, that are not affected by W_i. Let Y_i be the observed outcome of the ith unit given the assigned treatment, and $\{Y_i(1), \dots, Y_i(Z)\}$ the potential outcomes for the ith unit under each treatment of $\mathcal{W}$. For each unit, at most one of the potential outcomes is observed (the one corresponding to the treatment to which the unit is exposed). All other potential outcomes are missing, which is known as the fundamental problem of causal inference.[45] Let $r(w, \boldsymbol{X}_i)$ be the GPS, which is defined as the probability of receiving treatment w given pre-treatment covariates, that is $r(w, \boldsymbol{X}_i) = Pr(W_i = w|\boldsymbol{X}_i)$, for $\forall w \in \{1, \dots, Z\}$.[19,20] This definition extends the propensity score[21] from a binary treatment setting to the multiple treatment setting, in which conditioning must be done on a vector of GPSs, defined as $\boldsymbol{R}(\boldsymbol{X}_i) = (r(1, \mathbf{X}_i), \dots, r(Z, \boldsymbol{X}_i))$, or a function of $\boldsymbol{R}(\boldsymbol{X}_i)$.[20] In addition, we define the response surface as $f(w, \boldsymbol{X}_i) \equiv E[Y_i(w)|\boldsymbol{X}_i]$, for $w \in \{1, \dots, Z\}$.

In general, causal effects are not identifiable without further assumptions because only one of the potential outcomes is observed for every unit. We make the following identifying assumptions:

1. The stable unit treatment value assumption,[46] that is no interference between units and no different versions of a treatment.
2. The positivity or sufficient overlap assumption, that is $0 < p(W_i|Y_i(1), \dots, Y_i(Z), \boldsymbol{X}_i) < 1, \forall\, W_i \in \{1, \dots, Z\}$, which implies that there are no values of pre-treatment covariates that could occur only among units receiving one of the treatments.
3. The treatment assignment is unconfounded, that is $p(W_i|Y_i(1), \dots, Y_i(Z), \boldsymbol{X}_i) = p(W_i|\boldsymbol{X}_i), \forall\, W_i \in \{1, \dots, Z\}$, which implies that the set of observed pre-treatment covariates, $\boldsymbol{X}_i$, is sufficiently rich such that it includes all variables directly influencing both W_i and Y_i; in other words, there is no unmeasured confounding.

Under the unconfoundedness assumption, for any treatment w and pre-treatment covariates $\boldsymbol{X}_i$

$$f(w, \boldsymbol{X}_i) = E[Y_i(w)|W_i = w, \boldsymbol{X}_i] = E[Y_i|W_i = w, \boldsymbol{X}_i] \tag{1}$$

where the second identity is the conditional mean function of the observed outcomes.

2.2 Definition of causal effects

Causal effects are summarized by estimands, which are functions of the unit-level potential outcomes on a common set of units.[42,44] For dichotomous outcomes, causal estimands can be the RD, OD, or RR. For purposes of illustration, we use RD in this paper.

Following Lopez and Gutman,[14] we provide a broad definition of the causal RD that may be of interest with multiple treatments. Define s_1 and s_2 as two subgroups of treatments such that $s_1, s_2 \subset \mathcal{W}$ and $s_1 \cap s_2 = \varnothing$. Next, let $|s_1|$ and $|s_2|$ be the cardinality of s_1 and s_2, respectively. Two commonly used causal estimands are the ATE_{s_1,s_2}, and the ATE among those receiving s_1, $ATT_{s_1|s_1,s_2}$, where

$$\begin{aligned} ATE_{s_1,s_2} &= E\left[\frac{\sum_{w\in s_1} Y_i(w)}{|s_1|} - \frac{\sum_{w'\in s_2} Y_i(w')}{|s_2|}\right], \\ ATT_{s_1|s_1,s_2} &= E\left[\frac{\sum_{w\in s_1} Y_i(w)}{|s_1|} - \frac{\sum_{w'\in s_2} Y_i(w')}{|s_2|} | W_i \in s_1\right] \end{aligned} \tag{2}$$

In equation (2), the expectation is over all units, $i = 1, \ldots, N$, and the summation is over the potential outcomes of a specific unit. Another set of causal estimands are the conditional treatment effects, given pre-treatment covariates $\boldsymbol{X}_i$

$$\begin{aligned} CATE_{s_1,s_2} &= E\left[\frac{\sum_{w\in s_1} Y_i(w)}{|s_1|} - \frac{\sum_{w'\in s_2} Y_i(w')}{|s_2|} | \boldsymbol{X}_i; \theta_{Y|X}\right] \\ CATT_{s_1|s_1,s_2} &= E\left[\frac{\sum_{w\in s_1} Y_i(w)}{|s_1|} - \frac{\sum_{w'\in s_2} Y_i(w')}{|s_2|} | W_i \in s_1, \boldsymbol{X}_i; \theta_{Y|X}\right] \end{aligned} \tag{3}$$

where the parameters $\theta_{Y|X}$ and θ_X index the conditional distribution $p(Y|\mathbf{X})$ and distribution $p(\mathbf{X})$, respectively.

Causal inference methods via modeling the response surfaces (e.g. BART and RA) arrive at the population or sample marginal treatment effects by integrating the conditional effects over the distribution of $\boldsymbol{X}_i$.[47] In most cases, however, it is difficult to model the possibly multi-dimensional $\boldsymbol{X}_i$. We can obtain the marginal effects by averaging the treatment effects conditional on the observed values of the covariates over the empirical distribution of $\{\boldsymbol{X}_i\}_{i=1}^N$

$$\begin{aligned} ATE_{s_1,s_2} &= \int CATE_{s_1,s_2}(\boldsymbol{X}, \theta_{Y|X}) \mathrm{d}F_X(\boldsymbol{X}; \theta_X) \\ ATT_{s_1|s_1,s_2} &= \int CATT_{s1|s_1,s_2}(\boldsymbol{X}, \theta_{Y|X}) \mathrm{d}F_X(\boldsymbol{X}; \theta_X) \end{aligned} \tag{4}$$

In our motivating example, one of the research questions of interest is to compare the effectiveness of a newer minimally invasive procedure (i.e. robotic-assisted surgery) versus the existing surgical procedures (e.g. VATS) in the overall population, or among those patients who received robotic-assisted surgery. The corresponding target causal estimands are defined as

$$\begin{aligned} ATE_{1,2} &= \int E[Y_i(1) - Y_i(2) | \boldsymbol{X}_i; \theta_{Y|X}] \mathrm{d}F_X(\boldsymbol{X}; \theta_X) \\ ATT_{1|1,2} &= \int E[Y_i(1) - Y_i(2) | W_i = 1, \boldsymbol{X}_i; \theta_{Y|X}] \mathrm{d}F_X(\boldsymbol{X}; \theta_X) \end{aligned} \tag{5}$$

2.3 Treatment effects using BART

Under the identifying assumptions, treatment effects such as $ATT_{1|1,2}$ can be estimated by contrasting the imputed potential outcomes between robotic-assisted surgery and VATS groups among those patients who received robotic-assisted surgery, predicted from the estimates of the respective response surface models. In principle, any method that can flexibly estimate $f(w, \boldsymbol{X}_i)$ could be used to predict the potential outcomes. Chipman et al.[36,37] demonstrated that BART has important advantages as a predictive algorithm over alternative methods in the machine learning literature such as classification and regression trees,[48] boosting,[49] and random forests,[50] in particular with regard to choosing tuning parameters and generating coherent uncertainty intervals.

BART is a Bayesian ensemble method that models the mean outcome given predictors by a sum of trees. For a binary outcome, the BART model can be expressed using the probit model setup as

$$f(w, \boldsymbol{X}_i) = E(Y_i | W_i = w, \boldsymbol{X}_i) = \Phi\left\{\sum_{j=1}^{J} g_j(w, \boldsymbol{X}_i; T_j, M_j)\right\} \tag{6}$$

where Φ is the standard normal c.d.f., each (T_j, M_j) denotes a single subtree model in which T_j denotes the regression tree and M_j is a set of parameter values associated with the terminal nodes of the jth regression tree, $g_j(w, \boldsymbol{X}_i)$ represents the mean assigned to the node in the jth regression tree associated with covariate value $\boldsymbol{X}_i$ and treatment level w, and the number of regression trees J is considered to be fixed and known. The details of the specification of prior distribution and the choice of hyper-parameters can be found in Chipman et al.[37] Sampling from the posterior distributions proceeds via a Bayesian backfitting MCMC algorithm.[37] A total of L Markov Chain Monte Carlo (MCMC) samples of model parameters (T_j, M_j) are drawn from their posterior distribution. For each of L draws, we predict the potential outcomes for each unit and the relevant treatment level. The causal estimand of interest can be estimated by contrasting the imputed potential outcomes between treatment groups. For example, $ATT_{1|1,2}$ can be estimated as follows

$$\begin{aligned}\widehat{ATT}_{1|1,2} &= (n_1 L)^{-1} \sum_{l=1}^{L} \sum_{i:W_i=1} \left\{f^l(1, \boldsymbol{X}_i) - f^l(2, \boldsymbol{X}_i)\right\} \\ &= (n_1 L)^{-1} \sum_{l=1}^{L} \sum_{i:W_i=1} \left\{\Phi\left[\sum_{j=1}^{J} g_j(1, \boldsymbol{X}_i; T_j^l, M_j^l)\right] - \Phi\left[\sum_{j=1}^{J} g_j(2, \boldsymbol{X}_i; T_j^l, M_j^l)\right]\right\}\end{aligned} \tag{7}$$

where (T_j^l, M_j^l) are the lth draw from the posterior distribution of (T_j, M_j). We can obtain the point and interval estimates of the treatment effect directly using the summary of posterior samples.

2.4 Common support

Because problems can arise when drawing inference to regions of the covariate space where there are insufficient number of units in all treatment groups, propensity score-based methods are typically equipped with strategies for defining a common support region. For BART, there is no such a mechanism to prevent it from extrapolating over areas where a common support does not exist.

For a binary treatment, one strategy is to discard units that fall beyond the range of the propensity score.[51,52] Hill and Su[39] argue that these strategies typically ignore the information embedded in the response variable and propose alternative discarding rules. Illustrative examples with one or two predictors were used to compare the two types of discarding strategies and their implications on estimation of the causal effects and the proportion of inferential units retained. Advantages of BART over the propensity score approach manifest in examples where there is lack of common support for variables only predictive of treatment but not of the outcome or the treatment mechanism is more difficult to model. However, in practice, identifying common support is often required for a high-dimensional covariate space. In addition, the two types of strategies have not been compared in the multiple treatment setting.

To address these limitations, we propose a strategy for BART to define both a common support region and the corresponding discarding rules. Whereas Hill and Su[39] use a common support for binary treatment using the *1 sd rule*, our empirical simulations suggest this rule may be too relaxed in the setting of three or more treatment groups. We use a sharper cutoff and identify a common support as follows. We discard any unit i, with $W_i = w$, for which $s_i^{f_{w'}} > \max_j\{s_j^{f_w}\}$, $\forall j : W_j = w$, where $s_j^{f_w}$ and $s_j^{f_{w'}}$ denote the standard deviation of the posterior distribution of the potential outcomes under treatment $W = w$ and $W = w'$, respectively, for a given unit j.

For multiple treatments with $Z = 3$, when estimating the ATT of treatment $W = 2$ and $W = 3$ among those treated with $W = 1$, we discard for unit i with $W_i = 1$, if

$$\begin{aligned}&s_i^{f_2} > \max_j\{s_j^{f_1}\}, \text{ and} \\ &s_i^{f_3} > \max_j\{s_j^{f_1}\}\end{aligned} \tag{8}$$

When estimating the ATEs, we apply the discarding rule in equation (8) to each treatment group.

There is likewise a lack of consensus for defining a common support region with GPS-based approaches. For matching using the GPS, Lopez and Gutman[14] propose a rectangular support region. Let $r(w, \boldsymbol{X})$ denote the treatment assignment probability for w, and let $r(w, \boldsymbol{X}|W = w')$ represent treatment assignment probability for w among those who received treatment w'. A rectangular common support region can be defined as follows with $Z = 3$. For any $w, w' \in \mathcal{W} = \{1, 2, 3\}$

$$\begin{aligned} r(^{w,\boldsymbol{X})low} &= \max\{\min(r(w, \boldsymbol{X}|W = 1)), \min(r(w, \boldsymbol{X}|W = 2)), \min(r(w, \boldsymbol{X}|W = 3))\} \\ r(^{w,\boldsymbol{X})high} &= \min\{\max(r(w, \boldsymbol{X}|W = 1)), \max(r(w, \boldsymbol{X}|W = 2)), \max(r(w, \boldsymbol{X}|W = 3))\} \end{aligned} \tag{9}$$

For weighting methods, techniques such as trimming[32] or stabilizing (more useful for time-varying confounding, see Hu et al.,[53] Hu and Hogan,[54] and Hernán and Robins[55]) are frequently used in place of a common support. However, the lack of common support in the covariate space may lead to extreme weights and unstable IPTW estimators. In this article, we used fixed quantile-based trimming. More recently, Ju et al.[56] developed an adaptive truncation approach based on the collaborative TMLE methodology and showed their estimators achieved the best performance for both point estimation and confidence interval coverage among all propensity score truncation-based estimators.

3 Simulation studies

3.1 Design and implementation

We conduct expansive simulations in order to better understand how BART will work in complex causal settings. Our first set of simulations, Simulation 1, contrasts BART with other approaches, while our second set, Simulation 2, looks into the role that covariate overlap plays in inferences with multiple treatments.

The design of both simulations mimics the range of scenarios that are representative of the data structure in the SEER-Medicare registry. Three treatment levels ($Z = 3$) are used throughout, with pairwise ATTs of RD are our outcome of interest. True treatment effects are computed based on a simulated superpopulation of size 100,000. We replicated each of the scenarios described below 200 times within sub-populations of the superpopulation. In Simulation 1, we began with the comparisons of 10 methods: (1) RA, (2) IPTW with weights estimated using multinomial logistic regression (IPTW-MLR), (3) IPTW with weights estimated using generalized boosted models (IPTW-GBM), (4) IPTW with weights estimated using super learner (IPTW-SL), (5) IPTW-MLR with trimmed weights, (6) IPTW-GBM with trimmed weights, (7) IPTW-SL with trimmed weights, (8) VM, (9) TMLE, (10) BART. We used VM to only estimate the ATT effects as the algorithm for estimating the ATEs has not been fully developed, and implemented TMLE to only estimate the ATE effects as we are not aware of any implementation of TMLE for the estimation of ATT effects for multiple treatment options. In simulation 2, only BART, TMLE, and IPTW-GBM, the top performing methods in Simulation 1, were further examined.

We implemented the methods as follows. For RA, we first fit a Bayesian logistic regression model with main effects of all confounders using the bayesglm() function in the arm package in R. We then drew a total of 1000 MCMC samples of regression coefficients from their posterior distributions and predicted the potential outcomes for each unit and relevant treatment group. When implementing IPTW, we estimated GPSs by including each confounder additively to a multinomial logistic regression model, a GBM, and a SL model, respectively. The stopping rule for the optimal iteration of GBM was based on maximum of absolute standardized bias, which compares the distributions of the covariates between treatment groups.[24] We implemented SL using the weightit() function in the R package WeightIt for multinomial treatment and included three algorithms: main terms regression, generalized additive model, and support vector. The treatment probabilities are normalized to sum to one. The weights—inverse of the GPSs—were then trimmed at 5 and 95% to generate trimmed IPTW estimators. GPSs for VM were estimated using multinomial logistic regression with main effects of all confounders. We used a combination of k-means clustering with $k = 5$ subclasses and one-to-one matching with replacement and a caliper of 0.25 to ensure that the matched cohort is relatively similar in terms of the distributions of the confounders. We used the R package tmle to implement TMLE as described in Rose and Normand.[34] We used SL to estimate each treatment probability and bound them from below to 0.025. Applying BART to the simulation datasets, we used the default priors associated with the bart() function available in the BART package in R. For each BART fit, we allowed the maximum number of trees in the sum to be 100. To ensure the convergence of the MCMC in BART, we let the algorithm run for 5000 iterations with the first 3000 considered as burn-in.

To judge the appropriateness of each technique, we use mean absolute bias (MAB), root mean squared error (RMSE), and coverage probability (CP). In addition, we examine the large-sample convergence property of each method.

3.1.1 Simulation 1: Which causal approach yields the lowest bias and RMSE?

We compare each of the 10 approaches across a combination of two design factors: the study sample size (i.e. the total number of units) and the ratio of units in the treatment groups. We varied the two factors in three scenarios: (1) 1200 with a 1:1:1 ratio, (2) 4000 with a 1:5:4 ratio, and (3) 11,600 with a 1:15:13 ratio to represent equal, moderately unequal, and highly unequal sample sizes across treatment groups. The relatively small sample size (400) in the first group—which will be used as the reference group of the ATT effects—and the scenario of highly unequal sample sizes mimic the SEER-Medicare data in the motivating study.

We considered 10 confounders with five continuous variables and five categorical variables. We assumed that both the treatment assignment mechanism and the response surfaces are nonlinear models of the confounders, as a realistic representation of the application data. Specifically, the treatment assignment follows a multinomial logistic regression model

$$\begin{aligned} \ln\frac{P(W=1)}{P(W=3)} &= \alpha_1 + \mathbf{X}\xi_1^L + \mathbf{Q}\xi_1^{NL} \\ \ln\frac{P(W=2)}{P(W=3)} &= \alpha_2 + \mathbf{X}\xi_2^L + \mathbf{Q}\xi_2^{NL} \end{aligned} \tag{10}$$

where $\mathbf{Q}$ denotes the nonlinear transformations and higher-order terms of the predictors $\mathbf{X}$, ξ_1^L and ξ_2^L are vectors of coefficients for the untransformed versions of the predictors $\mathbf{X}$ and ξ_1^{NL} and ξ_2^{NL} for the transformed versions of the predictors captured in $\mathbf{Q}$. The intercepts, α_1, α_2, were specified to create the corresponding ratio of units in three treatment groups in each scenario. We generated three sets of parallel response surfaces as follows

$$\begin{aligned} E[Y(1)|\mathbf{X}] &= \text{logit}^{-1}\{\tau_1 + \mathbf{X}\gamma^L + \mathbf{Q}\gamma^{NL}\} \\ E[Y(2)|\mathbf{X}] &= \text{logit}^{-1}\{\tau_2 + \mathbf{X}\gamma^L + \mathbf{Q}\gamma^{NL}\} \\ E[Y(3)|\mathbf{X}] &= \text{logit}^{-1}\{\tau_3 + \mathbf{X}\gamma^L + \mathbf{Q}\gamma^{NL}\} \end{aligned} \tag{11}$$

where regression coefficients (τ_1, τ_2, τ_3, γ^L, and γ^{NL}) were chosen so that the prevalence rates in the treatment groups were similar as the rates of respiratory complications observed in the SEER-Medicare data (see Table 3). By generating nonparallel response surfaces across treatment groups, we can induce heterogeneous treatment effects. This topic warrants a stand-alone research and is beyond the scope of this article. Details of model specification in equations (10) and (11) are given in Table S1 of Supplementary Materials. The observed outcome Y is related to the potential outcome $Y(w)$ via $Y_i = \sum_{w\in\{w_1,w_2,\ldots,w_Z\}} Y_i(w)I(W_i = w)$.

3.1.2 Simulation 2: How do levels of covariate overlap impact causal estimates?

Only BART, TMLE, and IPTW-GBM, the top performing methods in Simulation 1, are used in Simulation 2, which more deeply examines the impact of covariate overlap.

We generated datasets following the simulation configuration of scenario 3 in Simulation 1, including the total sample size, the ratio of units, the number of continuous and categorical confounders, and the response surface models, to mimic the SEER-Medicare dataset. To create varying covariate overlaps that are "measurable" in degrees, we generate the treatment variable and covariate distribution as follows.

Three levels of covariate overlap were designed: (1) *weak*—there is lack of overlap in the covariate space defined by all 10 confounders, (2) *strong*—there is strong overlap with respect to each of the 10 confounders, and (3) *moderate*—the five categorical variables had sufficient overlap as in the *strong* scenario and overlap is lacking for the five continuous variables. Two configurations were examined in the *moderate* scenario. All of the five continuous variables or only two of them were included in the response surface models, resulting in one configuration where overlap was lacking for a variable that was a true confounder and another configuration when overlap was lacking for a variable that was not predictive of the response surface (therefore not a true confounder).

This simulation is designed to make it difficult for any method to successfully estimate the true treatment effect, as both the treatment assignment and the outcome are difficult to model. We simulated datasets for each scenario as follows.

- *Weak*. We assumed that the treatment variable W followed a multinomial distribution, $W \sim \text{Multinomial}(N, p_1, p_2, p_3)$, and generated the treatment assignment by setting $N = 11{,}600$, $p_1 = .03$, $p_2 = .52$, and $p_3 = .45$. The covariates were generated from the distributions conditional on treatment assignment to create sufficient or lack of overlap. The continuous variables were generated independently from $X_j|W = 1 \sim N(-1, 1)$, $X_j|W = 2 \sim N(1, 1)$, $X_j|W = 3 \sim N(3, 1)$ for $j = 1, \ldots, 5$. The categorical variables were generated independently from $X_j|W = 1 \sim \text{Multinomial}(N, .3, .3, .4)$, $X_j|W = 2 \sim \text{Multinomial}(N, .6, .2, .2)$, $X_j|W = 3 \sim \text{Multinomial}(N, .8, .1, .1)$, for $j = 6, \ldots, 10$. The potential outcomes of each treatment group were drawn from the response surface models (11), with all of the 10 covariates included (i.e. all covariates are true confounders). Under this scenario, lack of overlap was designed for each of the 10 confounders.
- *Strong*. The treatment variable W was generated in the same way as in the *weak* scenario. We created strong covariate overlap by generating similar distributions of the covariates across the treatment groups for all 10 confounders $X_1 - X_{10}$. Specifically, we assumed $X_j|W \sim N(.05W, 1 - 0.05W)$ for $j = 1, \ldots, 5$, and $X_j|W \sim \text{Multinomial}(N, .3 - .001W, .3 + .001W, .4)$ for $j = 6, \ldots, 10$.
- *Moderate*. We generated five categorical confounders $X_6 - X_{10}$ with strong overlap and lack of overlap for five continuous variables $X_1 - X_5$. We distinguished the situation where overlap is lacking for a variable that is not predictive of the outcome (*moderate* I) and the situation when it is lacking for a true confounder (*moderate* II). Specifically, we assumed that $X_j|W = 1 \sim N(-0.5, 1)$, $X_j|W = 2 \sim N(1, 1)$, $X_j|W = 3 \sim N(2.5, 1)$ for $j = 1, \ldots, 5$ and $X_j|W \sim \text{Multinomial}(N, .3 - .001W, .3 + .001W, .4)$ for $j = 6, \ldots, 10$. In *moderate* I, the response surface models only included covariates $X_1, X_5, X_6 - X_{10}$, thus, X_2, X_3, and X_4 that defined a covariate area in which the lack of overlap occurred are non-confounders. In *moderate* II, covariates $X_1 - X_{10}$ were all included in the response surface model, inducing lack of overlap in five true confounders.

Distributions of estimated GPSs across the treatments are compared using boxplots. For each overlap scenario, we estimated the GPS for each unit in the sample using GBMs, and plotted the distributions of estimated GPSs using a separate boxplot for the unit receiving each type of treatment (Figure 1). Substantial overlap in boxplots is presented in the strong overlap scenario, while the weak overlap scenario highlights the different distributions of GPSs.

3.2 Simulation results

3.2.1 Simulation 1

Table 1 presents the MAB, RMSE, and CP of the estimates of two ATT effects $ATT_{1|1,2}$ and $ATT_{1|1,3}$, and three ATE effects $ATE_{1,2}$, $ATE_{1,3}$, and $ATE_{2,3}$, for the three scenarios in Simulation 1.

No single method trumped others in estimating both $ATT_{1|1,2}$ and $ATT_{1|1,3}$ across all three scenarios. For $ATT_{1|1,2}$, outcome modeling approaches had smaller MABs and RMSEs, whereas for $ATT_{1|1,3}$, GPS approaches showed similar or slightly better performance than BART. RA performed best under the scenario of equal sample sizes. As the sample sizes in the comparison groups grew relative to the reference group, BART generally produced low MAB and RMSE. With GPS approaches, IPTW-GBM outperformed IPTW-MLR, IPTW-MLR-Trim, IPTW-SL, and IPTW-SL-Trim in the estimates of $ATT_{1|1,2}$ across all three scenarios, but had similar performances in estimating $ATT_{1|1,3}$. Weight trimming did not improve IPTW-MLR, IPTW-GBM, or IPTW-SL. VM presented larger bias and RMSE than BART and IPTW-GBM. None of the methods had nominal CP. IPTW methods and RA in general generated greater than the nominal CP, VM had a CP that decreased as the ratio of units became more unbalanced (0.99 to 0.80), and BART yielded a CP around 0.80–0.88, which we suspect is because the reference group is relatively small. Overall, BART and IPTW-GBM tended to show the best performances across settings for the ATT estimates.

For the ATE estimates, BART consistently provided lower MAB and RMSE followed by TMLE, across all three scenarios with different ratio of units. BART had nominal CP across all three scenarios. IPTW methods and TMLE yielded conservative intervals and greater than the nominal CP. RA was sensitive to the ratio of units. In the scenario with highly unequal sample sizes across treatment groups, RA had subpar performance. The intervals produced by RA rarely covered the true effects, resulting in a low CP. Altogether, BART and TMLE provided the

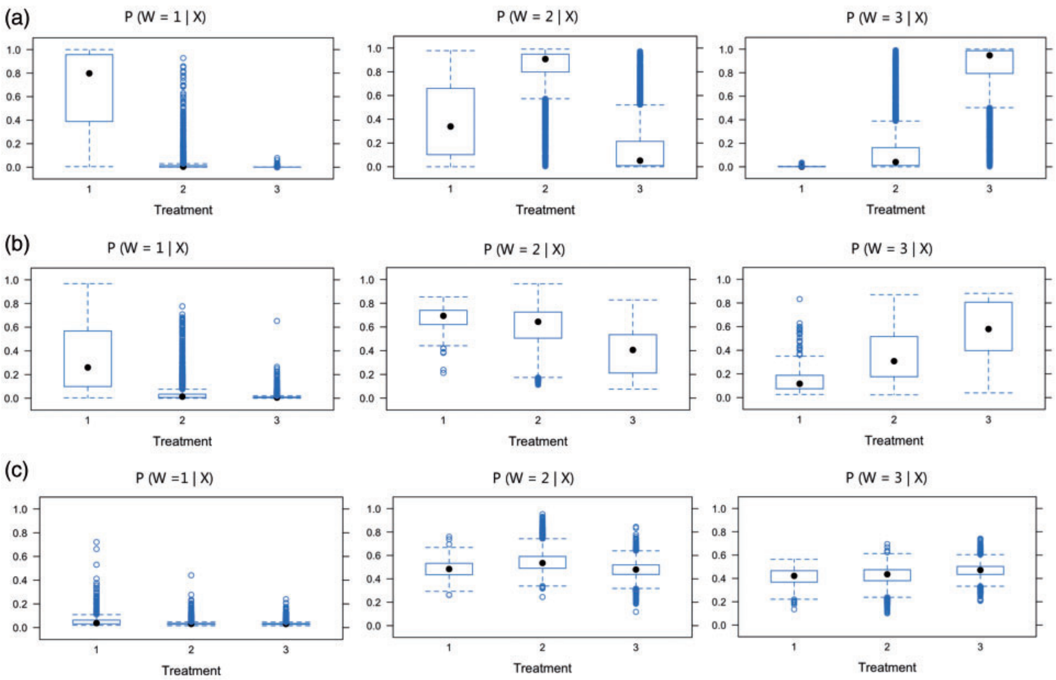

Figure 1. Overlap assessment for the scenarios of (a) weak, (b) moderate, and (c) strong covariate overlap. Each panel presents boxplots by treatment group of the estimated GPSs for one of the treatments, $P(W_i = w|X)$, $w \in \{1, 2, 3\}$, for every unit in the sample. The left panel presents treatment 1 ($W = 1$), the middle panel presents treatment 2 ($W = 2$), and the right panel presents treatment 3 ($W = 3$).

best performances across settings for the ATE estimates. Boxplots of biases from 200 replications in pairwise ATT and ATE estimates appear in Figures S1 and S2 in Supplemental Materials.

In Figure 2, we examined the large-sample convergence property of each of six methods. We considered only the scenario with the ratio of units = 1:15:13, which is the most representative of the SEER-Medicare registry. We simulated the data with increasing sample sizes of $n =$ (2900, 5800, 8700, 11,600, 14,500, 17,400). We computed the RMSE of the estimates of $ATT_{1|1,2}$ and $ATT_{1|1,3}$ for each n. We then regressed log(RMSE) on ($-$log$n$) using a simple linear regression with a slope b for each method. The least-squares estimation of b approximates the convergence rate.[57] BART and GBMs converged at a rate of $O(n^{-1/2})$ for both ATT estimates. IPTW-MLR, IPTW-SL, VM, and RA all converged at a slower rate than $O(n^{-1/2})$. Figure S3 in Supplemental Materials displays the convergence property of each of six methods for the estimates of the ATE estimates. BART and TMLE converged at a rate of $O(n^{-1/2})$ for all of the pairwise ATE estimates. GBMs varied in the rate of convergence across three pairwise ATE effects, from $O(n^{-1/2})$ to $O(n^{-2/5})$ to $O(n^{-1/3})$. IPTW-MLR, IPTW-S, and RA all had a much slower convergence rate.

3.2.2 Simulation 2

Figure 3 displays boxplots of biases of $ATT_{1|1,2}$ and $ATT_{1|1,3}$ among 200 simulations under four levels of overlap for each of IPTW-GBM, IPTW-GBM with trimmed weights, BART, and BART with discarding rules (Figure 3 (a)); and boxplots of biases of $ATE_{1,2}$, $ATE_{1,3}$, and $ATE_{2,3}$ for each of TMLE, BART, and BART with discarding rules (Figure 3(b)).

BART boasted smaller bias under nearly all levels of overlap compared to TMLE and IPTW-GBM. The advantage is more evident when there is more lack of covariate overlap. The larger biases and RMSEs (see Table S2 and Table S3 in Supplemental Materials) in the IPTW-GBM estimates under the *weak* scenario relative to *moderate* and *strong* overlap suggest that weighting by the GPS—even by employing flexible machine learning techniques—suffers from insufficient covariate overlap. The doubly robust method, TMLE, did not show as much variation in its performance across different levels of covariate overlap. In addition, in the *weak* scenario, weight trimming largely altered the IPTW-GBM estimates, indicating the lack of overlap may have led to extreme GPSs. GPS methods ignore the information in the outcome variable, thus assessing covariate overlap regardless of

Table 1. Comparison of the estimated average treatment effects on the treated in terms of mean absolute bias (MAB), root mean square error (RMSE), and coverage probability (CP) across 200 replications in Simulation 1. The causal estimand is based on risk difference.

		$ATT_{1\|1,2}$			$ATT_{1\|1,3}$			$ATE_{1,2}$			$ATE_{1,3}$			$ATE_{2,3}$		
Scenario	Method	MAB	RMSE	CP	MAB	RMSE	CP	MAB	RMSE	CP	MAB	RMSE	CP	MAB	RMSE	CP
I	RA	0.01	0.02	0.99	0.03	0.04	0.99	0.01	0.02	0.99	0.03	0.04	0.98	0.02	0.02	0.99
	IPTW-MLR	0.06	0.07	1	0.04	0.05	1	0.07	0.08	1	0.04	0.05	1	0.09	0.10	1
	IPTW-MLR-Trim	0.06	0.07	1	0.04	0.05	1	0.07	0.08	1	0.04	0.05	1	0.09	0.10	1
	IPTW-GBM	0.05	0.06	0.99	0.05	0.06	0.98	0.07	0.07	1	0.06	0.07	0.98	0.13	0.13	0.96
	IPTW-GBM-Trim	0.06	0.07	0.99	0.05	0.06	0.98	0.06	0.06	0.98	0.06	0.07	1	0.11	0.12	0.96
	IPTW-SL	0.06	0.07	1	0.05	0.06	1	0.07	0.08	1	0.05	0.06	1	0.12	0.13	1
	IPTW-SL-Trim	0.06	0.07	1	0.06	0.08	1	0.06	0.07	1	0.05	0.05	1	0.10	0.11	1
	VM	0.05	0.07	0.99	0.06	0.08	0.93	–	–	–	–	–	–	–	–	–
	BART	0.03	0.04	0.88	0.04	0.05	0.80	0.03	0.03	0.96	0.03	0.03	0.95	0.03	0.04	0.95
	TMLE	–	–	–	–	–	–	0.04	0.05	1	0.02	0.03	1	0.05	0.06	1
II	RA	0.02	0.02	1	0.05	0.05	0.92	0.02	0.02	0.80	0.05	0.05	0.60	0.03	0.03	0.67
	IPTW-MLR	0.05	0.06	1	0.03	0.03	0.99	0.06	0.08	1	0.04	0.05	1	0.07	0.07	1
	IPTW-MLR-Trim	0.06	0.06	1	0.03	0.03	0.99	0.06	0.07	1	0.03	0.04	1	0.08	0.08	1
	IPTW-GBM	0.03	0.04	0.98	0.03	0.04	0.99	0.05	0.05	0.98	0.05	0.06	1	0.09	0.09	0.94
	IPTW-GBM-Trim	0.05	0.06	0.98	0.04	0.04	0.99	0.05	0.05	0.98	0.05	0.06	1	0.09	0.09	1
	IPTW-SL	0.06	0.06	1	0.04	0.04	0.99	0.06	0.07	1	0.05	0.05	1	0.11	0.11	1
	IPTW-SL-Trim	0.06	0.07	1	0.06	0.06	0.99	0.06	0.07	1	0.05	0.05	1	0.10	0.10	1
	VM	0.04	0.05	0.86	0.05	0.07	0.88	–	–	–	–	–	–	–	–	–
	BART	0.02	0.03	0.80	0.03	0.04	0.75	0.02	0.02	0.96	0.01	0.02	0.98	0.01	0.02	0.94
	TMLE	–	–	–	–	–	–	0.04	0.04	1	0.02	0.02	1	0.04	0.04	0.96
III	RA	0.03	0.03	1	0.07	0.07	0.44	0.03	0.03	0.06	0.07	0.07	0.03	0.04	0.04	0.03
	IPTW-MLR	0.06	0.06	1	0.02	0.03	0.73	0.07	0.08	1	0.05	0.06	1	0.07	0.07	1
	IPTW-MLR-Trim	0.06	0.07	1	0.02	0.03	1	0.06	0.07	1	0.03	0.04	1	0.07	0.08	1
	IPTW-GBM	0.03	0.04	1	0.02	0.03	0.98	0.04	0.05	0.98	0.04	0.05	1	0.06	0.06	0.98
	IPTW-GBM-Trim	0.06	0.06	1	0.02	0.03	0.98	0.04	0.05	1	0.05	0.05	1	0.06	0.06	0.10
	IPTW-SL	0.06	0.06	0.99	0.03	0.03	1	0.06	0.07	1	0.04	0.05	1	0.10	0.10	1
	IPTW-SL-Trim	0.06	0.07	0.99	0.04	0.05	1	0.06	0.07	1	0.04	0.05	1	0.10	0.10	0.99
	VM	0.03	0.04	0.80	0.05	0.06	0.78	–	–	–	–	–	–	–	–	–
	BART	0.02	0.03	0.76	0.03	0.04	0.74	0.02	0.03	0.95	0.02	0.03	0.96	0.01	0.01	0.94
	TMLE	–	–	–	–	–	–	0.03	0.03	0.98	0.01	0.02	0.97	0.03	0.03	0.96

ATT: Average treatment effect on the treated; BART: Bayesian additive regression trees; IPTW-GBM: IPTW with weights estimated using generalized boosted models; IPTW-MLR: IPTW with weights estimated using multinomial logistic regression; IPTW-SL: IPTW with weights estimated using super learner; RA: regression adjustment; TMLE: Targeted maximum likelihood estimation; VM: vector matching.

whether the variables are true confounders; BART, on the contrary, takes advantage of the information contributed by the outcome. This is demonstrated by the similar performance delivered by IPTW-GBM in *moderate*I (lack of overlap in non-confounders) and *moderate*II (lack of overlap in true confounders), and better performance of BART in *moderate*I than in *moderate*II. BART perhaps recognized, in *moderate*I, that X_2, X_3, and X_4 do not play an important role in the response surface and showed a better performance than IPTW-GBM (smaller bias in both treatment effects).

Our BART discarding rule (8) considerably reduced the biases in the estimates of both ATE and ATT effects in the *weak* scenario where there was substantial lack of covariate overlap. When the lack of covariate overlap was moderate, the discarding strategy noticeably improved over plain BART. When there was sufficient covariate overlap, BART with and without discarding performed equally well. The weighting methods and TMLE are not coupled with discarding rules. To get a sense of the proportion of units that would be retained in the common support region for inference based on the GPSs, we applied the GPS-based discarding rule, employed by VM, designed for obtaining a common support region for multiple treatments (9). Using BART, the percentages of discarded units in the treated group, averaged across 200 replications, in the *weak*, *moderate*I, *moderate*II, and *strong* scenario were 38, 24, 15, and 0.2%, respectively, as compared to 86, 42, 42, and 13% computed by the GPS-based discarding rule (9). BART retains a much larger common support region while providing more accurate treatment effect estimates.

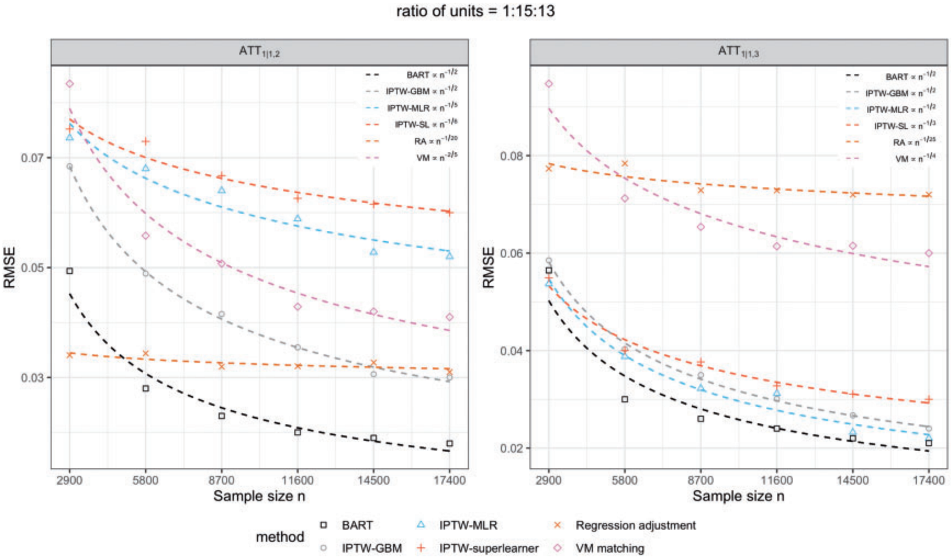

Figure 2. The large-sample convergence rate of each of six methods for the estimates of two treatment effects, $ATT_{1|1,2}$ and $ATT_{1|1,3}$. BART and IPTW-GBM converged the fastest, approximately at a rate of $O(n^{-1/2})$. RA converged the slowest, approximately at a rate of $O(n^{-1/20})$. ATT: Average treatment effect on the treated; BART: Bayesian additive regression trees; IPTW-GBM; IPTW with weights estimated using generalized boosted models; IPTW-MLR: IPTW with weights estimated using multinomial logistic regression; IPTW-SL: IPTW with weights estimated using super learner; RA: regression adjustment; RMSE: root mean squared error; VM: vector matching.

4 Application to SEER-Medicare data on NSCLC

Clinical encounter and Medicare claims data on 11,980 patients with stage I–IIIA NSCLC were drawn from the latest SEER-Medicare database. These patients were above 65 years of age, diagnosed between 2008 (first year patients in the registry underwent robotic-assisted surgery) and 2013, and underwent surgical resection via one of the three approaches, including robotic-assisted surgery, VATS, or open thoracotomy. The dataset contains individual-level information at baseline on the following variables: age, gender, marital status, race, ethnicity, income level, comorbidities, cancer stage, tumor size, tumor site, cancer histology, and whether they underwent positron emission tomography, chest computed tomography, or mediastinoscopy. Table 2 summarizes these variables for each surgical approach. We compared the effectiveness of the three surgical approaches in terms of four outcomes: the presence of respiratory complication within 30 days of surgery or during the hospitalization in which the primary surgical procedure was performed, prolonged length of stay (LOS) (i.e. > 14 days), intensive care unit (ICU) stay following surgery, and readmission within 30 days of surgery. Table 3 presents the outcome rates in the three surgical groups.

Among the 11,980 patients, 396 (3.3%) received robotic-assisted surgery, 6582 (54.9%) underwent VATS, and 5002 (41.8%) were operated via open thoracotomy. We estimated the causal effects of robotic-assisted surgery versus VATS or open thoracotomy among patients who underwent robotic-assisted surgery (i.e. $ATT_{s_1|s_1,s_2}$ and $ATT_{s_1|s_1,s_3}$) and in the overall population (i.e. ATE_{s_1,s_2} and ATE_{s_1,s_3}) using BART, RA, IPTW with GPSs estimated using multinomial logistic regression or GBMs (with or without trimming), and VM. Each method was implemented as described in Section 3. All pre-treatment covariates were included additively to the GPS models for IPTW methods and VM, and to the response surface models for RA and BART.

Table S4 in Supplemental Materials presents the point and interval estimates of $ATT_{s_1|s_1,s_2}$ and $ATT_{s_1|s_1,s_3}$ based on RD for all the methods examined. To provide uncertainty intervals for the treatment effect estimates, nonparametric bootstrap was used for the IPTW methods and VM, and Bayesian posterior intervals were used for RA and BART. All methods yielded statistically insignificant effects on respiratory complication and readmission if patients who received robotic-assisted surgery had instead been treated with open thoracotomy or VATS. For prolonged LOS and ICU stay, all methods except RA and VM suggested that robotic-assisted surgery led to significant smaller rates of the outcomes compared to open thoracotomy, but no statistically significant differences compared to VATS. The results from this empirical dataset provided partial evidence that robotic-assisted surgery

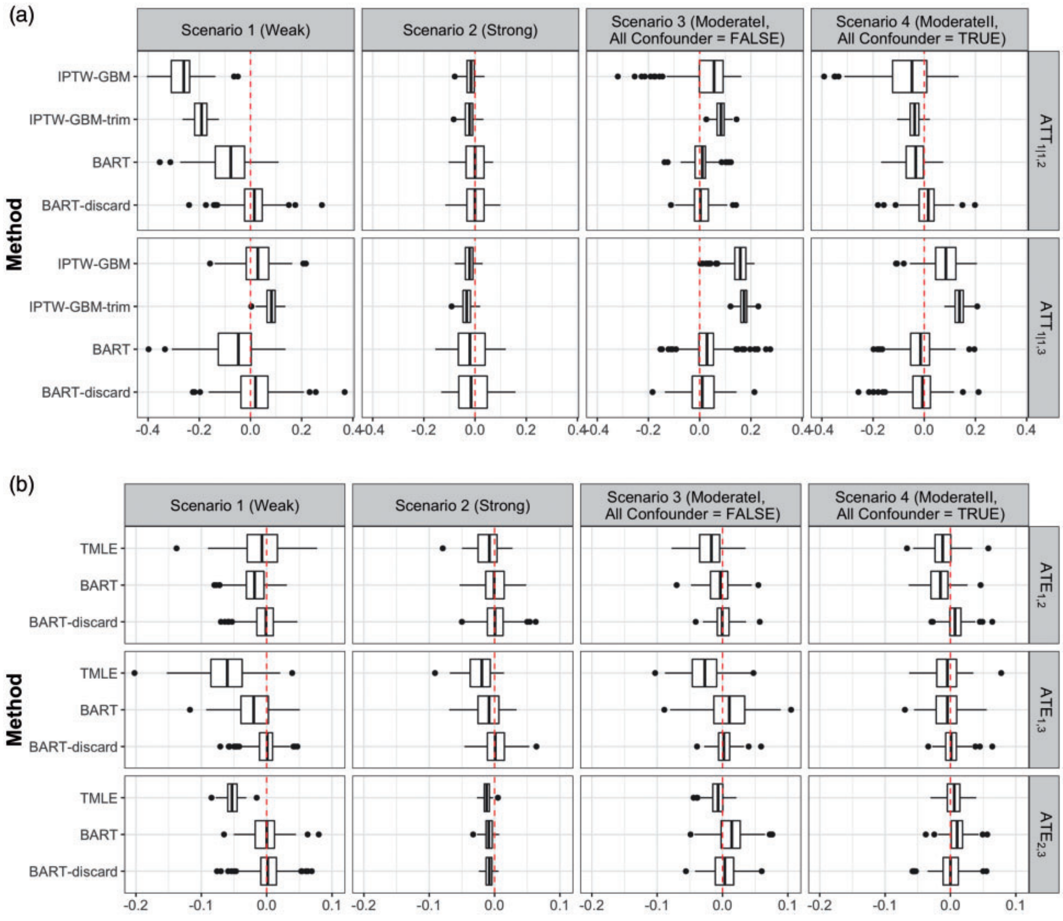

Figure 3. Biases among 200 replications under scenarios of differing covariate overlap for IPTW-GBM versus BART and two treatment effects $ATT_{1|1,2}$ and $ATT_{1|1,3}$; and for TMLE versus BART and three treatment effects $ATE_{1,2}$, $ATE_{1,3}$, and $ATE_{2,3}$. (a) BART-discard versus GBM for ATT estimates and (b) BART-discard versus TMLE for ATE estimates. ATE: Average treatment effects; ATT: Average treatment effect on the treated; BART: Bayesian additive regression trees; IPTW-GBM: IPTW with weights estimated using generalized boosted models; TMLE: targeted maximum likelihood estimation.

may have a positive effect on some postoperative outcomes among those who were operated with robotic-assisted surgery compared to open resection, but no advantages on over VATS.

To highlight the importance of simultaneous comparisons of multiple treatments, we implemented each method using SBCs to show how such inappropriate practices can result in different and confusing estimates of treatment effects. Table S4 also includes the estimates of $ATT_{s_1|s_1,s_2}$ and $ATT_{s_1|s_1,s_3}$ from SBCs. For BART, the conclusions are generally consistent with those using multiple treatment comparisons, though we note several inconsistent directions of the estimates of treatment effects. Given the different estimands and sub-populations to which inference using SBCs is generalizable when using GPS-based approaches, it would generally be inappropriate to directly compare causal estimates. However, we note that IPTW methods, implemented using SBCs, did not always match the findings that were based on IPTW methods designed for multiple treatments. The ATE estimates appear in Table S5 in Supplementary Materials.

We further explored the sensitivity of BART for binary outcomes to the choice of end-node prior, specifically via the hyperparameter k.[40] We employed 5-fold cross-validation to choose the optimal k that minimizes the misclassification error. Results suggested the optimal hyperparameter $k = 2$, which is the default value of k in the bart() function (not shown). Moreover, we extended the *1 sd rule*, the discarding rule of BART proposed by Hill and Su,[39] to the multiple treatment setting, to assess whether common support between treatment groups is reasonable based on the uncertainty in the posterior predictive distributions associated with the outcome in the observed versus the counter-factual treatment group. We did not exclude any patients from the empirical dataset based on the discarding rule in equation (8).

Table 2. Baseline characteristics of patients in three surgical groups in SEER-Medicare data.

Characteristics	Robotic-assisted surgery *N* = 396	VATS *N* = 6582	Open thoracotomy *N* = 5002
Age (years), mean (SD)	74.3 (5.7)	73.9 (5.4)	74.5 (5.7)
Female, N (%)	223 (56.3)	3446 (52.4)	2941 (58.8)
Married, N (%)	227 (57.3)	3753 (57.0)	2802 (56.0)
Race, N (%)			
White	320 (80.8)	5694 (86.5)	4369 (87.3)
Black	21 (5.3)	364 (5.5)	248 (5.0)
Hispanic	15 (3.8)	218 (3.3)	139 (2.8)
Other	40 (10.1)	306 (4.6)	246 (4.9)
Median household annual income, N (%)			
1st quartile	97 (24.5)	2132 (32.4)	1009 (20.2)
2nd quartile	88 (22.2)	1729 (26.3)	1193 (23.9)
3rd quartile	98 (24.7)	1345 (20.4)	1143 (22.9)
4th quartile	113 (28.5)	1376 (20.9)	1657 (33.1)
Charlson comorbidity score, N (%)			
0 − 1	154 (38.9)	2163 (32.9)	1810 (36.2)
1 − 2	113 (28.5)	1944 (29.5)	1379 (27.6)
>2	129 (32.6)	2475 (37.6)	1813 (36.2)
Year of diagnosis, N (%)			
2008–2009	14 (3.5)	2686 (40.8)	1484 (29.7)
2010	33 (8.3)	1123 (17.1)	857 (17.1)
2011	85 (21.5)	1033 (15.7)	866 (17.3)
2012	131 (33.1)	899 (13.7)	821 (16.4)
2013	133 (33.6)	841 (12.8)	974 (19.5)
Cancer stage, N (%)			
Stage I	295 (74.5)	4195 (63.7)	3884 (77.6)
Stage II	63 (15.9)	1504 (22.9)	709 (14.2)
Stage IIIA	38 (9.6)	883 (13.4)	409 (8.2)
Tumor size, in mm, N (%)			
≤20	160 (40.4)	1967 (29.9)	2232 (44.6)
21 − 30	98 (24.7)	1696 (25.8)	1388 (27.7)
31 − 50	109 (27.5)	1804 (27.4)	987 (19.7)
≥51	29 (7.3)	1084 (16.5)	367 (7.3)
Histology, N (%)			
Adenocarcinoma	255 (64.4)	3757 (57.1)	3348 (66.9)
Squamous cell carcinoma	107 (27.0)	2165 (32.9)	1167 (23.3)
Other histology	34 (8.6)	660 (10.0)	487 (9.7)
Tumor site, N (%)			
Upper lobe	215 (54.3)	3829 (58.2)	2859 (57.2)
Middle lobe	27 (6.8)	308 (4.7)	335 (6.7)
Lower lobe	141 (35.6)	2195 (33.3)	1720 (34.4)
Other site	13 (3.3)	250 (3.8)	88 (1.8)
PET scan, N (%)	302 (76.3)	5004 (76.0)	3410 (68.2)
Chest CT, N (%)	263 (66.4)	4525 (68.7)	3148 (62.9)
Mediastinoscopy, N (%)	62 (15.7)	715 (10.9)	420 (8.4)

CT: computed tomography; PET: positron emission tomography; SD: standard deviation; VATS: video-assisted thoracic surgery.

Table 3. The outcome rates in three surgical groups: robotic-assisted surgery, VATS, and open thoracotomy.

Outcomes	Robotic-assisted surgery *N* = 396	VATS *N* = 6582	Open thoracotomy *N* = 5002	Overall *N* = 11,960
Respiratory complication	30.1%	33.6%	33.3%	33.3%
Prolonged LOS	5.3%	10.4%	5.5%	8.2%
ICU stay	60.2%	75.3%	59.1%	67.9%
Readmission	8.8%	9.8%	8.0%	9.0%

ICU: intensive care unit; LOS: length of stay; VATS: video-assisted thoracic surgery.

5 Summary and discussion

Our paper makes two primary contributions to the causal inference literature. First, we extend BART to the multiple treatment and binary outcome setting, highlighting that the strengths of BART for binary treatment also manifest with multiple treatments. Second, we propose a common support rule for BART, and find that BART consistently shows superior performance over alternative approaches in various scenarios with differing levels of covariate overlap.

In addition to the primary findings in our simulations corresponding to bias, RMSE, CP, and large-sample convergence property, BART boasts a few additional advantages that make it a unique tool for the multiple treatment setting. As one example, BART is computationally efficient. All simulations were run in R on an iMAC with a 4 GHz Intel Core i7 processor. On a dataset of size $n = 11,600$, each BART implementation took less than 150 s to run, while each IPTW-GBM implementation took about 10 min to run. As a second example, BART produces coherent interval estimates of the treatment effects for either continuous or binary outcomes using posterior samples. For GBMs, McCaffrey et al.[24] estimate the variance by using robust procedure for continuous outcomes, but acknowledge that there is currently lack of theory to guarantee that this approach results in proper confidence intervals. For estimands based on a binary outcome such as the RD investigated in this article, it is difficult to approximate the variance using robust procedure. For matching-based approaches, there is still ambiguity regarding appropriate methods for interval estimation.[14,20,38]

We apply the methods examined to 11,980 stage I–IIIA NSCLC patients, drawn from the latest SEER-Medicare linkage. Results suggest that robotic-assisted surgery may be preferred in terms of prolonged LOS and ICU stay, among those who were operated via the robotic-assisted technology, relative to open thoracotomy or VATS. Different choice of methods, or inappropriate practice such as implementing SBCs for pairwise ATT effects, may lead to different conclusions about the treatment effects, explicating the importance of appropriate methods and practice for causal inference with multiple treatments.

The promising performance of BART in the complex multiple treatment setting will lay groundwork for several future research avenues. First, the flexibility offered by nonparametric modeling of BART can be leveraged to model regression relationships in survival data. Second, individual treatment effects that are easily obtained from BART provide a building block for estimating the heterogeneous treatment effect. Finally, we have made a significant untestable assumption related to unmeasured confounding. Developing sensitivity analyses under this complex multiple treatments setting leveraging BART would also be a worthwhile and important contribution.

Acknowledgements

This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the National Cancer Institute; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the SEER Program tumor registries in the creation of the SEER-Medicare database.

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 was supported in part by award ME_2017C3_9041 from the Patient-Centered Outcomes Research Institute and by grant P30CA196521-01 from the National Cancer Institute.

ORCID iD

Liangyuan Hu https://orcid.org/0000-0002-4067-892X

Supplemental material

Supplemental material for this article is available online.

References

1. Ferlay J, et al. *Global cancer observatory: cancer tomorrow*. Lyon: International Agency for Research on Cancer.
2. Molina JR, Yang P, Cassivi SD, et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc* 2008; **83**: 584–594.
3. Uramoto H and Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res* 2014; **3**: 242–249.
4. Scott WJ, Matteotti RS, Egleston BL, et al. A comparison of perioperative outcomes of video-assisted thoracic surgical (VATS) lobectomy with open thoracotomy and lobectomy: results of an analysis using propensity score based weighting. *Ann Surg Innov Res* 2010; **4**: 1.
5. Whitson BA, Groth SS, Duval SJ, et al. Surgery for early-stage non-small cell lung cancer: a systematic review of the video-assisted thoracoscopic surgery versus thoracotomy approaches to lobectomy. *Ann Thorac Surg* 2008; **86**: 2008–2018.
6. Park BJ, Melfi F, Mussi A, et al. Robotic lobectomy for non-small cell lung cancer (NSCLC): long-term oncologic results. *J Thorac Cardiovasc Surg* 2012; **143**: 383–389.
7. Wisnivesky JP, Smith CB, Packer S, et al. Survival and risk of adverse events in older patients receiving postoperative adjuvant chemotherapy for resected stages II-IIIA lung cancer: observational cohort study. *BMJ* 2011; **343**: d4013.
8. Yan TD, Black D, Bannon PG, et al. Systematic review and meta-analysis of randomized and nonrandomized trials on safety and efficacy of video-assisted thoracic surgery lobectomy for early-stage non-small-cell lung cancer. *J Clin Oncol* 2009; **27**: 2553–2562.
9. Toker A. Robotic thoracic surgery: from the perspectives of European chest surgeons. *J Thorac Dis* 2014; **6**: S211–S216.
10. Cajipe MD, Chu D, Bakaeen FG, et al. Video-assisted thoracoscopic lobectomy is associated with better perioperative outcomes than open lobectomy in a veteran population. *Am J Surg* 2012; **204**: 607–612.
11. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care* 2002; **40**: IV3–IV18.
12. Veluswamy RR, Whittaker SA, Nicastri DG, et al. Comparative effectiveness of robotic-assisted surgery for resectable lung cancer in older patients. *Am J Respir Crit Care Med* 2017; **195**: A4884.
13. Wisnivesky JP, Henschke CI, Swanson S, et al. Limited resection for the treatment of patients with stage IA lung cancer. *Ann Surg* 2010; **251**: 550–554.
14. Lopez MJ and Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. *Stat Sci* 2017; **32**: 432–454.
15. Agresti A. *Categorical data analysis*. Hoboken, NJ: John Wiley & Sons, 2003.
16. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007; **26**: 3078–3094.
17. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol* 2008; **61**: 537–545.
18. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med* 2010; **29**: 2137–2148.
19. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710.
20. Imai K and Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *J Am Stat Assoc* 2004; **99**: 854–866.
21. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
22. Linden A, Uysal SD, Ryan A, et al. Estimating causal effects for multivalued treatments: a comparison of approaches. *Stat Med* 2016; **35**: 534–552.
23. Feng P, Zhou XH, Zou QM, et al. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med* 2012; **31**: 681–697.
24. McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013; **32**: 3388–3414.
25. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 1973; **29**: 185–203.
26. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J Am Stat Assoc* 1979; **74**: 318–328.
27. Imbens GW and Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press, 2015.
28. Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952; **47**: 663–685.
29. Little RJ. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988; **6**: 287–296.
30. Kang JD, Schafer JL, et al. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; **22**: 523–539.
31. Cole SR and Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008; **168**: 656–664.

32. Lee BK, Lessler J and Stuart EA. Weight trimming and propensity score weighting. *PLoS One* 2011; **6**: e18174.
33. Van der Laan MJ, Polley EC and Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**: 1–21.
34. Rose S and Normand SL. Double robust estimation for multiple unordered treatments and clustered observations: evaluating drug-eluting coronary artery stents. *Biometrics* 2019; **75**: 289–296.
35. Schuler MS and Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol* 2017; **185**: 65–73.
36. Chipman HA, George EI and Mcculloch RE. Bayesian ensemble learning. In: Schölkopf B, Platt JC and Hoffman T (eds) *Advances in neural information processing systems 19*. Cambridge: MIT Press, 2007, pp.265–272.
37. Chipman HA, George EI, McCulloch RE, et al. BART: Bayesian additive regression trees. *Ann Appl Stat* 2010; **4**: 266–298.
38. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 2011; **20**: 217–240.
39. Hill J and Su YS. Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann Appl Stat* 2013; **7**: 1386–1420.
40. Dorie V, Harada M, Carnegie NB, et al. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat Med* 2016; **35**: 3453–3470.
41. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (1990). *Stat Sci* 1923; **5**: 465–472.
42. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
43. Rubin DB. Assignment to treatment group on the basis of a covariate. *J Educ Stat* 1977; **2**: 1–26.
44. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978; **6**: 34–58.
45. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
46. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc* 1980; **75**: 591–593.
47. Ding P and Li F. Causal inference: a missing data perspective. *Stat Sci* 2018; **33**: 214–237.
48. Breiman L, Friedman J, Stone CJ, et al. *Classification and regression trees*. Boca Raton, FL: CRC Press, 1984.
49. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; **55**: 119–139.
50. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
51. Dehejia RH and Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat* 2002; **84**: 151–161.
52. Morgan SL and Harding DJ. Matching estimators of causal effects: prospects and pitfalls in theory and practice. *Sociol Methods Res* 2006; **35**: 3–60.
53. Hu L, Hogan JW, Mwangi AW, et al. Modeling the causal effect of treatment initiation time on survival: application to HIV/TB co-infection. *Biometrics* 2018; **74**: 703–713.
54. Hu L and Hogan JW. Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics* 2019; **75**: 695–707.
55. Hernán MA and Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.
56. Ju C, Schwab J and van der Laan MJ. On adaptive propensity score truncation in causal inference. *Stat Methods Med Res* 2019; **28**: 1741–1760.
57. Liu T, Hogan JW, Wang L, et al. Optimal allocation of gold standard testing under constrained availability: application to assessment of HIV treatment failure. *J Am Stat Assoc* 2013; **108**: 1173–1188.