PATIENT-CENTERED OUTCOMES RESEARCH INSTITUTE

FINAL RESEARCH REPORT

New Statistical Methods to Assess How Patients with Different Traits Respond to the Same Treatment

Daniel J. Feaster, PhD; Hemant Ishwaran, PhD; Min Lu, PhD; Saad Sadiq, MS

AFFILIATION:

University of Miami School of Medicine, Miami, Florida

Institution Receiving Award: University of Miami School of Medicine
Original Project Title: Methods for Heterogeneity of Treatment Effects: Random Forest Counterfactual Machines
PCORI ID: ME-1403-12907
HSRProj ID: HSRP201512317

To cite this document, please use: Feaster DJ, Ishwaran H, Lu M, Sadiq S. (2020). *New Statistical Methods to Assess How Patients with Different Traits Respond to the Same Treatment*. Patient-Centered Outcomes Research Institute (PCORI). https://doi.org/10.25302/06.2020.ME.140312907

TABLE OF CONTENTS

ABSTRACT

Background: We developed statistical methods to estimate individual-specific treatment effects (TEs) that can be used to find patient, clinical, and contextual characteristics that define subgroups that are more or less likely to respond to a treatment. This research supports PCORI's call for the development of analytic approaches and guidance for predictive approaches to the heterogeneity of TEs (HTE).

Objectives: Aim 1 was to develop comprehensive methodology (a) to estimate individual TEs (ITEs) for patients using counterfactual random forest (RF) machines for heterogeneous data, which (b) work across differing magnitudes of HTE, when (c) data are observational with potential confounding, and (d) with survival outcomes and multiple treatment comparisons.

Aim 2 was to develop software to implement these new methods in a wide array of comparative effectiveness research applications by adding functionality to the existing CRAN-distributed randomForestSRC package developed by our group, a package in its fifth release at the beginning of this project.

Methods: Our approach to the estimation of ITEs builds on Rubin's causal model using a counterfactual approach and expands this approach from a focus on uncovering average TEs (ATEs) to use on ITEs. Aims 1a through 1c were each addressed with a series of simulations. Aim 1a compared RF estimations using a variation of the virtual twin (VT)[1] approach with both a multiple-imputation approach[2] and Bayesian additive regression trees (BART)[3] to estimate ITEs assuming random assignment to treatment. Various models for generating heterogeneity were tested with complex interaction effects. Aim 1b varied the amount or size of HTE, sample size, random error, and level of prognostic covariates (which predict outcome regardless of treatment). Aim 1c compared 7 different RF-based approaches, 2 that were VT based,[1] a bivariate RF-based imputation approach,[4] counterfactual RF,[5] counterfactual synthetic RF, causal RF,[6] and BART[3] on confounded data. In all cases, the results were examined conditional on the propensity score to show that methods performed well across the distribution of propensity (including regions where confounding was highest). Aim 1d was addressed with an application to an observational clinical cohort of patients with ischemic cardiomyopathy. This application compares 4 different treatments, explicitly models overlap in the patient population, and incorporates expert knowledge. This procedure first estimates a propensity score for treatment assignment and then uses a new distance-based RF algorithm that creates a measure of distance among the patients in the sample using an unsupervised RF algorithm. Calibration creates cut points for the inclusion of particular individuals in particular treatment comparisons.

Results: The results for aim 1a showed that both RF and BART performed well with no mean bias in ITE estimates, with BART having the smallest root mean square error (RMSE) across simulations. Multiple imputation showed substantial bias for 1 data generation model and had the highest RMSE across simulations. The results for aim 1b showed that as the size of both prognostic factors that predict outcome (but not differential outcome across treatments)

and/or random error increases, so does the length of the confidence intervals around ITE estimates. Similarly, as the amount of heterogeneity increases, so do the individual confidence intervals for ITE estimates. We identified an edge bias phenomenon, wherein RF ITE estimates have increasing bias toward the ATE as the true TE becomes more extreme (ie, further from the ATE). The results for aim 1c showed that RF procedures performed well with observational confounded data as long as the confounders were in the feature set. Further, we found that synthetic RF had the lowest bias and smallest RMSE of the methods compared. BART had nearly as good performance as that of synthetic RF. Finally, the results of aim 1d showed that RF methods can estimate ITE with survival outcomes and multiple treatment comparisons. With survival outcomes, the ITE is a difference in survival functions, leading to multiple points of comparison and potential reversals between early and later dominance of different treatments.

Conclusions: RF is a flexible method that can be used to estimate ITEs; however, there are situations in which other methods may outperform RF. Future research should examine ways of assessing when particular methods are optimal.

Limitations: RF estimators of ITEs have mean bias across the distribution of ITEs at or near zero in most data generation models considered; however, more work is needed to understand the uncertainty pertaining to these individual estimates.

BACKGROUND

There is substantial heterogeneity within a population in the way individuals respond to a specific treatment.[7-11] However, typical clinical trial procedures focus on whether a treatment on average is effective and frequently actively discourage the examination of subgroup interactions.[12-20] Patients would like to know what is going to work for them, but statistical procedures are focused on the average response.

To address this problem, our application developed statistical methods to find patient, clinical, and contextual characteristics that define subgroups that are more or less likely to respond to a treatment. As noted by Alexander and Lambert,[21] the potential promise of comparative effectiveness research (CER)— identifying the right treatment for the individual— is at risk if the focus is on comparing differences in average treatment effects (TEs) between competing treatments when there is actually heterogeneity of TEs (HTE) across individuals. Our proposed research is responsive to the PCORI recommendation to support the development of analytic approaches and guidance for predictive approaches to HTE. With respect to PCORI methodology standards,[22] the methods would be considered hypothesis generating, and we provide guidance on how to prespecify the analysis plan.

Both behavioral prevention[7] and medical[8-11] research have frequently shown prominent subgroup differences in TE. However, subgroup analyses of clinical trial data are controversial,[12-20,23] largely due to the high likelihood of type I error and lack of replicability[23] when an exhaustive list of interactions with treatment are tested. Machine learning techniques, such as random forests[24] (RFs), provide a principled approach to explore a large number of predictors and identify replicable sets of predictive factors. In recent innovations, these machine learning techniques have been used specifically to uncover subgroups with differential treatment responses.[1,11,25-30] Some of these, such as the virtual twin[1] (VT) approach for the identification of subgroups, build on the idea of counterfactuals, that is, the outcome that would have been observed if an individual was assigned to the treatment opposite of the one to which they truly were assigned.[31] The VT approach uses RF as a first step to create separate predictions of outcomes under both treatment and control conditions for each trial participant

by estimating the counterfactual treatment outcome. In the second step, tree-based predictors are used to uncover the features/variables that explain differences in the person-specific TE and the characteristics associated with subgroups. This promising procedure does tend to have relatively low sensitivity and positive predictive value. We evaluated an improvement on this approach by creating treatment-specific counterfactual machines. These procedures have been incorporated into a user-friendly, freely available R package for future use in CER.

The data application clinical area that we planned to use to develop these CER methods to address HTE was sexually transmitted infection (STI) and HIV. Project AWARE,[32] a large (N = 5012) randomized comparative effectiveness trial, found that HIV risk reduction counseling for HIV-negative individuals at the time of an HIV test did not have an impact on the cumulative incidence of STI. However, the question remains as to whether there are subgroups that would benefit from counseling. The primary outcome of Project AWARE was STI incidence, a binary categorical outcome; however, secondary outcomes included continuous and count outcomes, such as the total number of condom-less sexual episodes and number of partners.

Our survival model methods were applied to a secondary analysis of ischemic cardiomyopathy data that was previously reported.[48] These data include the 1468 patients who were treated for ischemic cardiomyopathy (defined as severe left ventricular systolic dysfunction with ejection fraction of less than 30%) at the Cleveland Clinic between 1997 and 2007. Data were obtained from a prospective registry approved by the University of Miami School of Medicine IRB with waiver of patient consent. Cohort patients were categorized into the following 4 different treatments: (1) coronary artery bypass graft (CABG, n = 386); (2) CABG plus mitral valve annuloplasty (MVA); (CABG+MVA, n = 212); (3) CABG plus surgical ventricular restoration (SVR) (CABG+SVR, n = 360); (4) and listing for cardiac transplantation (LCTx, n = 510). The outcome modeled was all-cause mortality, including postsurgical death as well as death while awaiting transplantation. Patients were followed with consent, and this was supplemented with a search of the Social Security Death Index with a closing date of February 16, 2007. This resulted in 5577 patient-years of follow-up, with 3.8 ± 2.8 years of follow-up on average. There were a total of 444 events (deaths) over the follow-up period with totals by

treatment of CABG (n = 186/386); CABG+SVR (n = 84/360); CABG+MVA (n = 80/212); and LCTx (n = 174/510).

Not all patients are potential candidates for each of the surgical procedures or LCTx. The methods for comparing treatment outcomes at the individual level need to account for this. CABG is a surgical procedure that was developed in the late 1960s for patients with severe coronary artery disease. CABG+MVA is generally only done if severe mitral valve regurgitation is present. CABG+SVR addresses ventricular dysfunction, and although it has been shown to reduce ventricular volume relative to CABG alone, it was not shown to reduce rates of death or hospitalization.[33] Finally, LCTx requires severe clinical symptoms and clinical doubt as to the appropriateness of the surgical interventions.

This clinical example highlights the need for methods to assess HTE in a systematic fashion. The clinical reality of many diseases like ischemic cardiomyopathy is that there are multiple treatment choices, and clinical knowledge leads to decisions to try to match patients to appropriate treatment. Having statistical methods to quantify these decisions would be an important aid to decision-making and could allow better targeting of interventions in the future.

This methodology will also innovate CER generally. PCORI has called for researchers to include "participants representative of the spectrum of the population of interest" and to "identify and assess participant subgroups" in their standards for research questions. However, as noted previously, the statistical and clinical trial community of scholars has repeatedly focused on the well-established pitfalls of subgroup analysis without sufficient concern for its benefits.[12-20,23] Therefore, the concern about subgroup analyses and concerns of internal validity in efficacy/effectiveness research and resultant homogeneous samples have prevented the accumulation of trial-based knowledge about HTE. There are many clinical examples of heterogeneity of response derived from trial-based evidence, mostly from meta-analyses. However, meta-analyses even when examining heterogeneity may miss important subgroups.[34] The increasing emphasis on CER creates an opportunity to study HTE systematically at the patient level, if a replicable method of identifying patient subgroups with differential response

were available. The present application aimed to develop, test, and provide software to implement an innovative strategy for HTE.

Original Aims

Our original application had 3 aims.

Aim 1: Develop comprehensive methodology, which will

(1a) estimate patient TEs using person-specific counterfactual RF (CF) machines for heterogeneous and potentially confounded observational data;

(1b) identify subgroups of patients with differential TEs using a novel unsupervised RF;

(1c) compare differential TE identification against a novel bump hunting algorithm; and

(1d) derive theoretical properties and run extensive empirical tests and synthetic simulations to assess the efficacy of the proposed methods.

Aim 2: Develop software to implement these new methods in a wide array of CER applications by developing a user-friendly R package with API to add to the CRAN R package randomForestSRC (https://cran.r-project.org/web/packages/randomForestSRC/index.html) developed by our group, a package in its fifth release.

Aim 3: Work with stakeholders (ie, public health department and infectious disease clinicians) and patient groups (ie, men who have sex with men [MSM] and other patients at the STI clinic) in all phases of research, in particular with interpretation of the clinical and personal value of the model as implemented on the Project AWARE data. Part of this aim was also to explore the feasibility of incorporating data collection and predictive modeling as part of clinic procedures.

In the course of doing this research, it became apparent that some of these proposed aims needed to be altered. In the case of aims 1b and 1c, both unsupervised RF and bump

hunting are trying to find discrete subgroups with different characteristics and outcomes. Thus, they depend on empirical clustering in the data. We found, in Project AWARE, that our approach created individual treatment estimates that were quite smooth and normal in distribution, resulting in a continuum of heterogeneity with little evidence of clumping or clustering into categorical subgroups. Therefore, it was not necessary to develop methods to find clustering but rather to focus on issues affecting the individual treatment effects (ITEs). Second, it became clear that this research was at too early a stage for meaningful stakeholder engagement activities. Therefore, working with our PCORI project officer and contract managers, we modified aims 1 and 2 as follows.

Modified Aims

Aim 1 was to develop comprehensive methodology to

(1a) estimate ITEs using person-specific CF machines for heterogeneous data;

(1b) work across differing magnitudes of HTE;

(1c) work when data are observational with potential confounding; and

(1d) work with survival outcomes and with multiple treatment comparisons.

Aim 2 was to develop software to implement these new methods in a wide array of CER applications by adding functionality to the existing CRAN-distributed randomForestSRC package developed by our group, a package in its fifth release at the beginning of this project.

PARTICIPATION OF PATIENTS AND OTHER STAKEHOLDERS

Our original intent was to engage both patients and stakeholders, as is clear in our original aim 3.

Throughout the study, we did meet with an advisory board of stakeholders that included representatives of the public health department, several infectious disease doctors, and a community researcher with close ties to the target population. They found the approach interesting and potentially promising, but it also became clear that implementation with patients within the 3 years of the contract was not realistic.

Our original plan had been to conduct focus groups with patient groups to obtain preliminary data to plan a study where we would actually try to predict who might benefit from brief risk reduction counseling. Given that most of the heterogeneity in treatment was centered on a negative impact of treatment, we modified this contract and did not pursue the patient focus groups in the last year of the project.

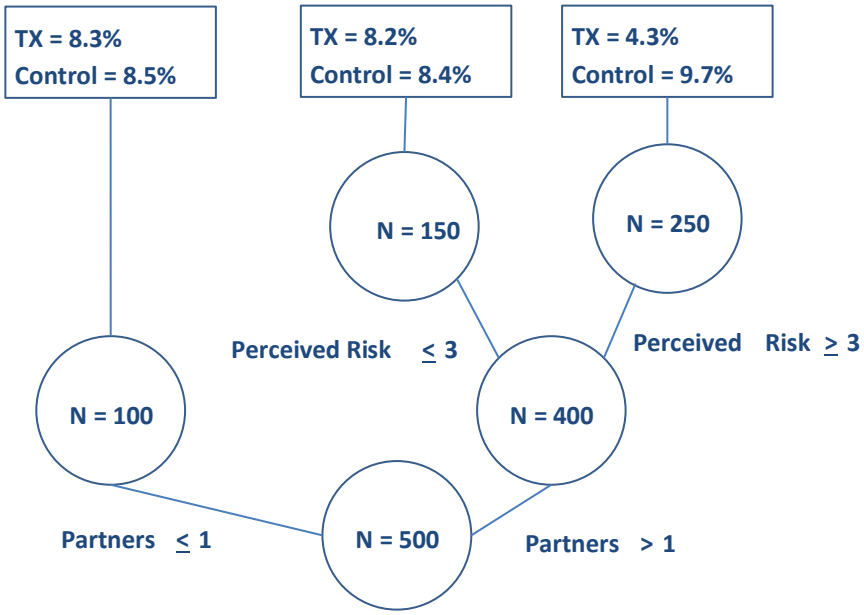
METHODS

Background on RF and Trees

Our methodology falls within the realm of statistical learning theory; sometimes, the language in this domain is unfamiliar even to individuals with considerable statistical background. We therefore begin with some definitions of various terms and techniques to which we will refer later. Our proposal uses an ensemble learning method called *random forest.*[24] Statistical learning techniques attempt to take simple inputs, such as observed characteristics, and build predictive models for some output or outputs that can be improved over time and with additional experience. In statistics, the inputs are frequently called predictors or independent variables, but in learning approaches, the term *feature* is used interchangeably with predictor. Similarly, in much of statistics, the outputs are called responses or dependent measures. The term *learner* is used to describe an elementary modeling technique, such as linear regression, which is used as the engine for learning or prediction. RF uses classification and regression tree (CART)[4] learners. CART is a nonparametric method (ie, a method free of model assumptions) that uses recursive partitioning to choose a series of cut points for maximizing differentiation between groups defined by the predictors.[35] This results in a series of nodes where the population can be divided and branches are as shown in the example in Figure 1. The process starts at the bottom of the tree (root node comprising the entire sample), with the root node split into 2, where the split point is determined by the value of the single predictor maximizing group differentiation. The groups in the resulting nodes are then split again by finding new nodes on which to split the subpopulations in those nodes. This process continues recursively until it is no longer possible to identify groups that differ on the outcome or the sample size at that node is too small, at which point a terminal node (top of the tree) has been reached. Figure 1 displays a hypothetical regression tree using estimated TE in risk reduction counseling as the outcome (see equation 1 for what we mean by an estimated TE). After comparing all potential splits on all variables, CART determines that having ≤1 or >1 sexual partner was the split in the data that explained the largest (initial) variability in the TE. It found no additional splits below the ≤1 partner node. It did find an additional split for >1 partner node at ≤3 on the perceived risk scale. The results show that patients in the right-most

terminal node, with >1 sex partner and with perceived risk for HIV >3, had improved STI incidence in the treatment group. The strength of CART is that it provides a general approach for predicting outcomes; predictions do not rely on model assumptions; nonlinear relationships are accommodated; and multiway interactions can be accommodated in the prediction. However, a well-known drawback of CART is its instability, that is, small changes in the data result in markedly different trees, which leads to high variability.[36] RF is a state-of-the-art ensemble learning method designed to address the instability of CART.[24,36] Ensemble learners can be loosely defined as predictors formed by aggregating base learners. In RF, CART is used for the learner; however, multiple trees are grown, each on different bootstrapped samples of the sample. In total, a collection of *ntree* >1 random trees are grown, which are aggregated to form the ensemble predicted value. Also, during the tree-growing process, at each node of the tree, a random subset of variables is chosen of size $1 \leq mtry \leq p$, where p equals the total number of independent variables. The node is split using the variable from the *mtry* candidate variables yielding the best split. Splitting is repeated recursively as in CART, with the tree grown as far as possible, while maintaining the condition that each terminal node contain a minimum of *nodesize* ≥1 unique individuals in the sample. A small value of *nodesize* is typically used in order to induce deep trees with many nodes and branches (the original description of RF[24] used splitting to purity in classification problems, ie, *nodesize* = 1). Once the forest is grown, each person's set of independent variables can be used to find what terminal node they are in (or would have been in if they were not part of the bootstrapped sample for a particular tree). The average of responses in that terminal node is the predicted value for that tree for all individuals in that particular terminal node. The predicted value for an individual is calculated by averaging their individual tree predictions across all trees in the forest.

Figure 1. A Simple CART Diagram



Abbreviations: CART, classification and regression tree; TX, treatment.

The superior performance of RF to CART can be attributed to the randomization involved in growing its trees and the low biased nature of its learners. The 2-step randomization of growing a tree using bootstrapped data, and the use of random feature selection, has the effect of decorrelating the different trees in the forest, which yields a low variance predictor. Meanwhile, by using deeply grown trees, low bias is simultaneously achieved. Thus, RF achieves both low bias and low variance, unlike CART, which generally can only achieve 1 factor at the price of the other. In general, RF has proven to be an excellent predictor, performing especially well in challenging problems, including scenarios when the number of variables exceeds the sample size (the so-called "*p* bigger than *n*" problem), in settings where complex interactions are at play between variables, and when predictors are nonlinear. These properties have enabled RF to be successfully applied in many scientific problems.[37-49]

Synthetic RFs

We will also be using a new approach to RFs: synthetic RFs. RFs have tuning parameters, *nodesize* and *mtry*, the random number of features to be considered at each node, which can be chosen to maximize prediction or minimize a loss function, such as mean square error (MSE). This involves estimating many RFs with different values of the tuning parameters and choosing

the forest that minimizes or maximizes your objective. An alternative approach is to use each of the forests from the procedure to create a predicted value for the outcome of interest under each of the various *nodesize* and *mtry* specifications. These predicted values are called synthetic features because they are synthesized from the data. These synthetic features are then appended to the original features and a final forest estimated on the original features and the synthetic features, jointly. This approach is called synthetic RF.[50] We will be examining both the original approach to RFs and the synthetic RF approach.

RF Counterfactual Machines

To describe our approach to aim 1a, we begin by introducing some notation. Let $\{(T_1,\mathbf{X}_1,Y_1),...,(T_n,\mathbf{X}_n,Y_n)\}$ denote the data where $\mathbf{X}_i$ is the p-dimensional covariate (feature, independent variable) for patient i and Y_i is the outcome. We assume that Y_i is a binary outcome, $Y_i \in \{0,1\}$, but our methodology applies to general outcomes; for example, it applies to multiclass (categorical) outcomes $Y_i \in \{C_1,...,C_J\}$ and continuous outcomes. Variables T_i record the treatment for patient i. For simplicity, we assume for the moment that the treatment is 1 of 2 values, $T_i \in \{0,1\}$. Each patient i is administered 1 of the 2 treatments. Thus, $T_i = 0$ or $T_i = 1$; however, we would like to know what the predicted probability for Y_i is under both treatment regimens, even though we know very well that the patient can only experience 1 treatment. That is, we would like to predict $p_{i,0} = P\{Y_i = 1|T_i = 0,\mathbf{X}_i,\}$ and $p_{i,1} = P\{Y_i = 1|T_i = 1,\mathbf{X}_i\}$.

To do so, we create 2 RF machines under each treatment type and use 1 for counterfactual inference. An RF machine for treatment $T = 0$ is constructed by running RF classification using only the data for patients with treatment $T_i = 0$. We call this machine RF_0. Likewise, an RF machine for treatment $T = 1$, denoted by RF_1, is constructed by running RF classification using only data with $T_i = 1$.

Now, given a patient, i, with $T_i = 0$, we obtain i's predicted value, $\hat{p}_{i,0}$, using RF_0. To obtain i's counterfactual probability, we assume there is a clone of patient i that is identical to i in all ways except that the clone has received the alternate treatment. Thus, the clone has an identical $\mathbf{X}_i$ (p-dimensional covariate) but differs because $T_i = 1$. Then, to obtain the clone's

estimated outcome, we simply drop the clone down RF_1 (ie, we apply the rules of RF_1 to the clone's features) and obtain the predicted probability, $\hat{p}_{i,1}$, which represents *i*'s counterfactual probability estimate. The value

$$Y_i^* = \left(\hat{p}_{i,0} - \hat{p}_{i,1}\right) \qquad (1)$$

represents the estimated TE for patient *i* (when $T_i = 1$, a TE estimate is obtained in an analogous fashion using RF_0 as the counterfactual machine). By performing this operation on each person in the sample, we can obtain a prediction of each individual's expected TE. These machines could also be used to create predictions for new individuals to inform treatment decisions.

Extension to More Than 2 Treatments

The approach can be extended to the setting where there are more than 2 treatment options. Say, for example, there are *M* treatment types, denoted by $T \in \{1,...,M\}$. Calculate RF machines by stratifying on each treatment. This yields *M* machines $RF_1,...,RF_M$. For each *i,* calculate *i*'s predicted probability and its *M* − 1 counterfactual predicted probabilities from its *M* − 1 counterfactual machines. Denote the resulting predicted probabilities by $\hat{p}_{i,1}, \ldots, \hat{p}_{i,M}$. Let $\hat{p}_i = \sum_{j=1}^{M} \hat{p}_{i,j}/M$ be the overall treatment mean. Define the TE for treatment *j* as

$$\hat{Y}_i^j = \left(\hat{p}_{i,j} - \hat{p}_i\right). \quad 1 \le j \le M$$

There are *M* distinct TEs. To each of these, we apply our algorithm.

Cross-validated Estimates

To further improve reliability and stability of inference, we exclusively use cross-validated estimates obtained using "out-of-bag" (OOB) prediction. Recall that RF trees are calculated using bootstrap samples. On average, 36.7% of the data are excluded from any bootstrap sample because these people happen not to be included by the random choice of individuals. This OOB data can be used to calculate cross-validated estimates. For example, the

OOB predicted value for patient *i* is calculated using only those trees for which *i* is OOB. Because this estimate does not involve *i*'s data, it represents a valid cross-validated estimate. As another example, when calculating the proximity between patients *i* and *j*, OOB proximity will be used, this being calculated by using only those trees for which *i* and *j* are both OOB.

Confounding

Each treatment may be to a degree bounded within constraints of indication and appropriateness. Certain treatments may simply not be suitable for certain patients. Counterfactual probabilities calculated without adjusting for such confounding will be biased. When calculating counterfactual probabilities, we first ascertain that the data used for the counterfactual machine are balanced for the patient; otherwise, that machine cannot be used for estimating a TE for that patient. One popular method to address this confounding in making comparisons of outcomes of alternative treatments is propensity score analysis.[51] Rubin characterizes an observational cohort study as a broken randomized trial, and the propensity score is the key to finding the mechanism for treatment selection to approximate a randomized clinical trial.[52] Propensity score analysis involves 2 steps. The first calculates the probability of receiving 1 or another treatment as a function of confounding variables. Assume a setting involving 2 treatment types, that is, $T \in \{0,1\}$. As a first step, one fits a propensity model by logistic regression of treatment T on the covariate $\mathbf{X}$. The conditional probability of receiving the intervention given $\mathbf{X}$ is the propensity score, denoted here by $P(\mathbf{X})$:

$$P(\mathbf{X}) = \mathrm{P}\{T = 1|\mathbf{X}\}. \qquad (2)$$

The propensity score has the balancing property such that T and $\mathbf{X}$ are conditionally independent given $P(\mathbf{X})$. The second step in a classical propensity analysis is to match individuals on their propensity scores; individuals in each treatment group are included in the analysis only if there is a matching individual with the same propensity in the other treatment group. The balancing property shows that for patients with the same propensity score, their confounders have the same distributions, regardless of treatment group. Therefore, the confounders are balanced between the 2 comparative groups after matching, and the matching process approximates a randomized clinical trial.[52] Note of course that propensity matching is

not required in our approach, as the patient's clone is the matched patient (this is because the clone has an identical **X** and therefore an identical propensity score $P(\mathbf{X})$). However, we still need to ascertain that the data on which the counterfactual machine was trained include the values of the clone's **X**.

Propensity score analysis was developed for 2-group comparisons and does not apply to examples involving multiple groups. Recently, the methodology was extend to more general treatment regimens that include multiple treatment groups and a theoretical framework that extended the notion of a propensity score to that of a generalized propensity function, and its associated generalized propensity score (GPS) was described.[53] Theoretical development of the GPS for multiple groups has shown that when the theoretical propensity score is modeled, say, as a multinomial probability model, the resulting (vector) propensities are balancing scores, and that TE estimates that depend on the propensity function are unbiased and not affected by confounding. However, there remains a gap between the GPS theory and practice, as the theory demands flexible, data-driven approaches that work with complex data while[53] historically relying on classical parametric models using maximum likelihood estimation. Instead, we propose using an RF machine for estimating the GPS. Indeed, the idea of using RF for propensity analysis has been successfully explored recently in the literature, although only for the 2-group propensity score analysis.[54] Very promisingly, it was found that the propensity score from RF resulted in better balance and bias reduction than did logistic regression.[54] The following theorem gives a unified definition of the GPS and forms the basis of our approach.

Theorem 1: *Suppose there exist functions g and P such that* $T = {}^{d} g(P(\mathbf{X}),\varepsilon)$*, where ε is a random vector that is independent of* **X***. Then, T and* **X** *are conditionally independent given* $P = P(\mathbf{X})$*. Call* $P(\mathbf{X})$ *the GPS.*

Theorem 1 says that if we can decompose the treatment *T* into 2 parts, 1 of which is completely explained by **X** and the other being completely random, then the first part must be the GPS. Importantly, this result is fully nonparametric and does not make any a priori assumptions about the relationship between *T* and **X**. For example, it does not assume a logistic relationship as is commonly adopted in propensity score analysis. The following corollary shows

how to find $P(\mathbf{X})$. The result is merely a restatement of the well-known probability integral transform.

Corollary 1: *Let $F_{T|\mathbf{X}}(t)$ be the right-continuous cumulative distribution function of the* conditional *distribution of T given* **X**. *Then, T can be written as* $T \overset{d}{=} F^{-1}_{T|\mathbf{X}}(U)$, *where U is independent of* **X** *and U has a uniform distribution. That is, the GPS must be $P(\mathbf{X}) = F_{T|\mathbf{X}}(t)$.*

For example, when *T* is a binary treatment, corollary 1 shows that the GPS is nothing more than the conditional probability of treatment given **X** (cf Lamont et al[2]). On the other hand, when *T* is a treatment with *M* treatment types, the GPS is a vector of $M - 1$ conditional multinomial probabilities. These values are directly calculated in an RF analysis and therefore allow us to immediately calculate the GPS. The multinomial probability RF analysis is based on the following steps:

1. Fit an RF analysis using treatment *T* as the outcome and **X** as the covariates. Determine the GPS, $P(\mathbf{X}_i)$, for each patient *i*.

2. To determine balancedness, the estimated $\hat{P}(X_i)$ should be examined to ensure that each element is bounded away from 0 and 1.

Unsupervised RF for Subgroup Identification

Unsupervised RF applies when there is no outcome response. The available data are *X* variables, which are either continuous or discrete (categorical), but there is no *Y* variable. We address this type of unsupervised learning problem using the following modified RF approach. In place of the standard *mtry* random feature selection, we instead randomly select *mtry* subsets of *q*-tuples of variables at each tree node where $1 \le q < p$. For each *q*-tuple, 1 variable from the *q*-tuple is chosen at random to be the feature to be split on, and the remaining $q - 1$ variables are treated as the responses (called the pseudo-outcomes). The best split (measured in terms of the mixed-outcome splitting rule) over the *mtry q*-tuples is used to split the node.

When $q = 2$, the splitting rule used is either weighted MSE splitting when the pseudo-outcome is continuous or Gini splitting when the pseudo-outcome is categorical. Weighted MSE

and the Gini splitting rule are well-known CART splitting rules. Recently, we have shown they are members of the same class of splitting rules and therefore share similar theoretical properties.[55] Thus, it is perfectly valid to compare the split statistic value from a regression pseudo-outcome split to the split statistic value from a categorical pseudo-outcome split. This only requires that the continuous pseudo-outcomes be rescaled by standardizing by their variance in the parent node, which is simple and computationally efficient to implement.

When $q > 2$, the pseudo-outcome is multivariate with dimension $q - 1$ and is used when **X** is expected to be highly correlated, as splitting on multiple coordinates will improve subgroup identification. To split multivariate outcomes, we use a composite splitting rule, defined as the average of the splitting rule applied to each coordinate. For coordinates that are continuous, weighted MSE splitting is used, whereas Gini splitting is used for categorical variables. Outcomes that are continuous are standardized by their variance. Doing so ensures that the coordinate average split statistic values are valid due to the theoretical equivalence of Gini and weighted splitting.[55]

The unsupervised forest returns a proximity matrix P of dimension $n \times n$, where entry (i,j) reflects the closeness of patient i to j. To determine subgroups from this, we convert the proximity matrix to a distance matrix $D = 1 - P$ and then apply clustering to D; for example, we will use hierarchical clustering with a target of K distinct clusters. The K distinct clusters represent our K distinct subgroups. A summary patient is created to characterize each subgroup. The summary patient is used to create a measure of distance for each patient as described in the split-merge algorithm.

RF Algorithm for Subgroup TE Identification

Below, we outline the RF algorithm used to identify subgroups of patients with TEs. Various details, such as the split-merge algorithm and variable selection, are presented in subsequent sections.

1. Calculate RF machines RF_0 and RF_1.
2. Calculate estimated TEs $\tilde{Y}_1,\ldots,\tilde{Y}_n$ using (1).

3. Run an RF regression (RF-R) using $\tilde{Y}_1,...,\tilde{Y}_n$ for the outcomes and $\mathbf{X}_1,...,\mathbf{X}_n$ as the covariates.

4. Apply the split-merge algorithm (see next section) to the predicted values from the RF-R of step 3.

5. This yields groups $g = 1,...,G$ with expected TE increasing in g. Within each group are K distinct subpopulations. Thus, groups with larger values of g will consist of 1 or several subpopulations of patients with sizable TEs that differ from the average TE.

Split-Merge Algorithm

A key step in the RF algorithm for subgroup identification is the split-merge RF algorithm that we now describe. Fundamentally, the split-merge algorithm relies on the notion of proximity, a key quantity calculated in an RF analysis. The forest proximity of an individual i is an n-dimensional vector $\mathbf{p}_i = (p_{i,1},\cdots,p_{i,n})$, whose jth entry, $p_{i,j}$, equals the forest relative frequency of i and j sharing the same terminal node (here, n is the sample size). This value measures the closeness of patient i to j. The split-merge algorithm involves the following steps (the algorithm assumes an RF model has been fit already):

1. Group patients into $G > 1$ groups corresponding to the percentile of their RF-predicted value of $\tilde{Y}_I$, their estimated TE, obtained from an RF model (eg, group patients in increments of 5 percentiles of $\tilde{Y}_1$). A large value of G is used to induce refined groups. A latter step automatically merges groups based on forest proximity; thus, a large value of G is not problematic and, in fact, is beneficial.

2. For each group g, decide whether to leave each patient (called the target patient) within its group g or move the patient to an adjacent group (for the 2 end groups, there is only 1 adjacent group, $g + 1$ or $g - 1$; for all others, there are 2 adjacent groups, $g - 1$ and $g + 1$). To make this determination, calculate a representative patient for each of the adjacent groups and the initial group. The representative patient is the "summary" patient within the group, defined by forming a p-dimensional vector composed of a summary value of each of the p variables. For discrete variables, the summary is a majority vote, where the category with the most people in it is the summary value; otherwise, the summary is calculated by taking the average. The summary patient that is closest in forest proximity defines the group to which the target patient is moved (or if the selected summary patient is from the initial group, the target patient is not moved).

Patients are assessed for movement across the groups sequentially starting with the highest predicted value.

3. To improve the potential for finding a good match within each adjacent group and within the initial group, identify $K \geq 1$ distinct subgroups using unsupervised RF. In total, there are K distinct summary patients within an adjacent group and K distinct summary patients within the initial group who are used to determine proximity to the target patient. These subgroups are subject to a minimum membership restriction of typically 5 or 10 patients.

4. Apply steps 2 and 3 in turn to each group $g = 1,...,G$, each time updating group membership after completing a step g.

5. At completion, the algorithm returns a new grouping of patients. Using this as an initial starting value, the algorithm is repeated. This process can be repeated several times; however, in our experience, the algorithm converges rapidly (2 to 3 iterations, but 1 iteration is also sufficient).

Applying steps 1 to 5 yields a classification of each patient into 1 of G groups. To remove instability in ranking, the entire split-merge algorithm is bootstrapped. More precisely, a bootstrap sample of the data is drawn, and steps 1 to 5 are applied to these data. This is repeated B times independently. For each patient, this results in a set of B values $\{\hat{G}_{i,1},...,\hat{G}_{i,B}\}$, where $\hat{G}_{i,b} \in \{1,...,G\}$ is the group assignment for patient i for bootstrap b. The average $\hat{G}_i = \sum_{b=1}^{B} \hat{G}_{i,b}/B$ (rounded) represents the final group assignment for i. The set of values $\{G_1,...,G_n\}$ is used to classify the data into G groups. Within each group g, we apply unsupervised RF to obtain K subgroups. The K subgroups within g represent K subpopulations that have similar outcome behavior but with different patient characteristics.

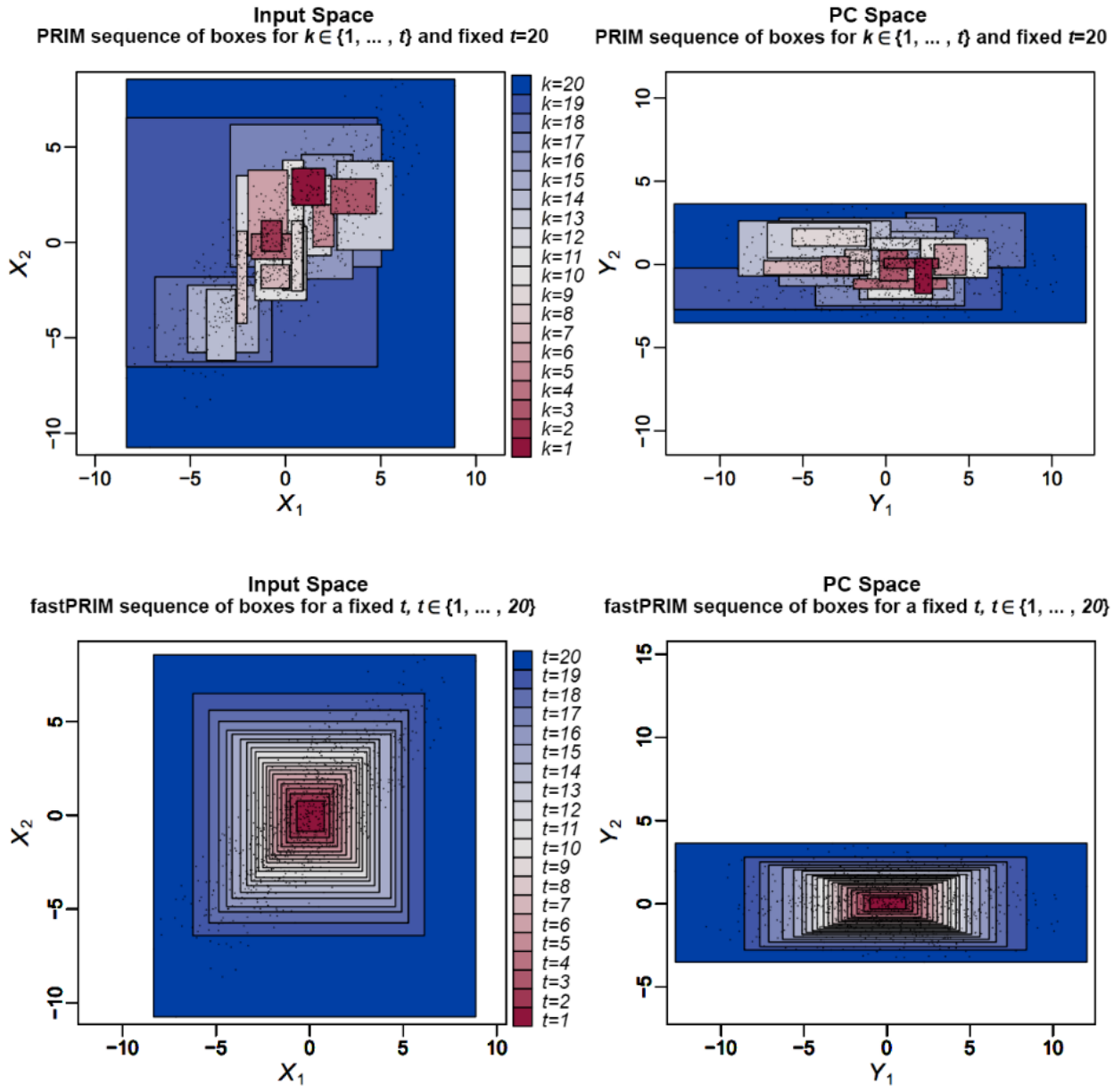
Bump Hunting for Subgroup Identification

We had originally planned to explore an alternate strategy for identifying important subgroups with potentially different TEs. The patient rule induction method (PRIM) algorithm for bump hunting was first developed by Friedman and Fisher.[56] It is an intuitively useful computational approach for the detection of local maxima (ie, bumps) on target functions (Figure 2). Its objective is to find subregions in the input space with relatively high values for the response variable. By construction, PRIM targets these subregions directly rather than indirectly

through estimation of a regression function through a series of *peeling* and *pasting* steps acting along the directions of the predictor variable coordinate axes. The method is such that these subregions can be described by simple rules, such as the union of rectangles in the input space. The PRIM algorithm can, however, perform poorly with correlated predictors and when the number of predictors becomes large. Dazard and Rao[57] noticed that PRIM could be significantly improved by incorporating the joint structure in the covariate space via a principal component analysis (PCA) transformation (ie, PRIM-PCA). More specifically, they proposed a local sparse bump hunting (LSBH) strategy that divides the predictor space into subregions where at most 1 node is present, then a sparse PCA transformation (SPCA) is performed separately on each local region with a node, and, finally, the location of the bump is determined via PRIM in the local, rotated, and projected region induced by the SPCA. This strategy was used effectively to identify heterogeneous subgroups with respect to survival in patients with colon cancer.[57,58] The LSBH algorithm is, however, computationally expensive, so Diaz et al[59] proposed modified versions of both PRIM and PRIM-PCA called *fastPRIM* that takes advantage of situations where symmetric predictor distributions exist by using a very efficient peeling and pasting algorithm that can greatly reduce the number of steps required. The fastPRIM algorithm with PCA rotation was also shown to possess certain optimality properties, including the fact that it produces boxes with minimum volume, thus providing a more accurate characterization (with respect to the predictor space) of the rectangular box approximating the bump. We will use the LSBH and fastPRIM algorithms as an alternate strategy for identifying important subgroups. To visualize the difference between these approaches, consider Figure 2, which is an illustration of simulated data where a bump is hiding in 2-dimensional predictor space. It is visually very clear what the differences in the approaches are algorithmically and how different the resulting bumps can look. The first row of plots depicts PRIM operating on the nonrotated (left) and PCA-rotated (right) predictor spaces. The second row of plots is the same but this time for fastPRIM. Notice how the PCA rotation produces a more focused search along the direction of 1 of the 2 predictors, that with higher variance. Notice also that the symmetric peeling and pasting of fastPRIM are clearly evident, resulting in distinctly different final subregions (the darkest-red

rectangles) and that fastPRIM in the PCA space produces the most “concentrated” dark-red rectangle.

Figure 2. PRIM and fastPRIM Ssequences of Bboxes as a Ffunction of Ccoverage[a]



Abbreviations: PC, principal component; PRIM, patient rule induction method.

[a]Top row: PRIM complete sequence of peeled boxes. Bottom row: fastPRIM complete series of boxes. Results are given in the input space (left) and in the PC space (right). The red-to-blue palette corresponds to a range of box output means from the largest to the smallest, respectively.

Methods for Individual Simulation Projects

Methods of Studies Associated With Aims 1a Through 1d

For aims 1a through 1c, we devised a series of simulation studies. In these studies, we generated data in which we knew the true TE for each individual for each treatment to which they could be assigned. We then tested our procedure(s) with 250 to 1000 different generated data sets to understand how the procedure would normally work. We had 3 major simulation studies in this project and 1 empirical example. The first 2 simulations focused on showing that our proposed methods would work on clinical trial data. An earlier project related to these 2 sets of simulations was initiated before this PCORI project. We also describe the results of this earlier project because it frames how we thought about these 2 simulations. Study 1 of the PCORI simulations attempted highly nonlinear data generation models and compared how different approaches to estimation worked when treatment was randomly assigned. Study 2 examined how ITE estimates performed as the amount of heterogeneity and the sample size on which the RF counterfactual machines were estimated varied, again when treatment was randomly assigned. Study 3 examined how ITE estimates were affected when confounded observational data were used to estimate each of the machines. Finally, study 4 illustrated how to estimate ITEs on survival outcomes and when there were multiple treatments available, some of which might not be available to an individual. This final study also examined how to incorporate expert knowledge into rankings of treatment outcomes.

Our initial work showing that RF can estimate individual treatment outcomes. We have completed extensive work in this area, including 1 paper that has been published[2] (note that because most of the work on this manuscript was completed before the PCORI project was funded, this has not been counted as one of the products of this contract) and 1 paper that is close to completion. The published manuscript compared 2 methods of generating predictions of how individuals would react to treatments, RF, and multiple imputation (MI). The MI approach assumes that the relationship between individual characteristics and individual outcomes from treatment follows a multivariate normal distribution. As noted in the overview, RF makes very few assumptions. It is known as a

nonparametric method and does not assume anything about the statistical distribution of individual characteristics, the treatment outcome, or the functional form that links an individual's characteristics to their likely treatment outcome.

MI procedure to generate ITEs. In general, our approach to generating person-specific TEs uses what is called a potential outcomes framework. In this framework, each individual has a potential response under every treatment condition. Of course, there is only 1 actual response (ie, the response under the treatment the individual actually receives; responses under other treatment conditions are not observed). Although we observe people in both treatments, we do not observe the same person in both treatments. Conceptualized in this way, the unobserved values associated with the unobserved treatment conditions could be considered a missing data problem and handled with modern missing data techniques. In a randomized clinical trial, the missing data are completely due to randomization and are therefore known to be missing completely at random.

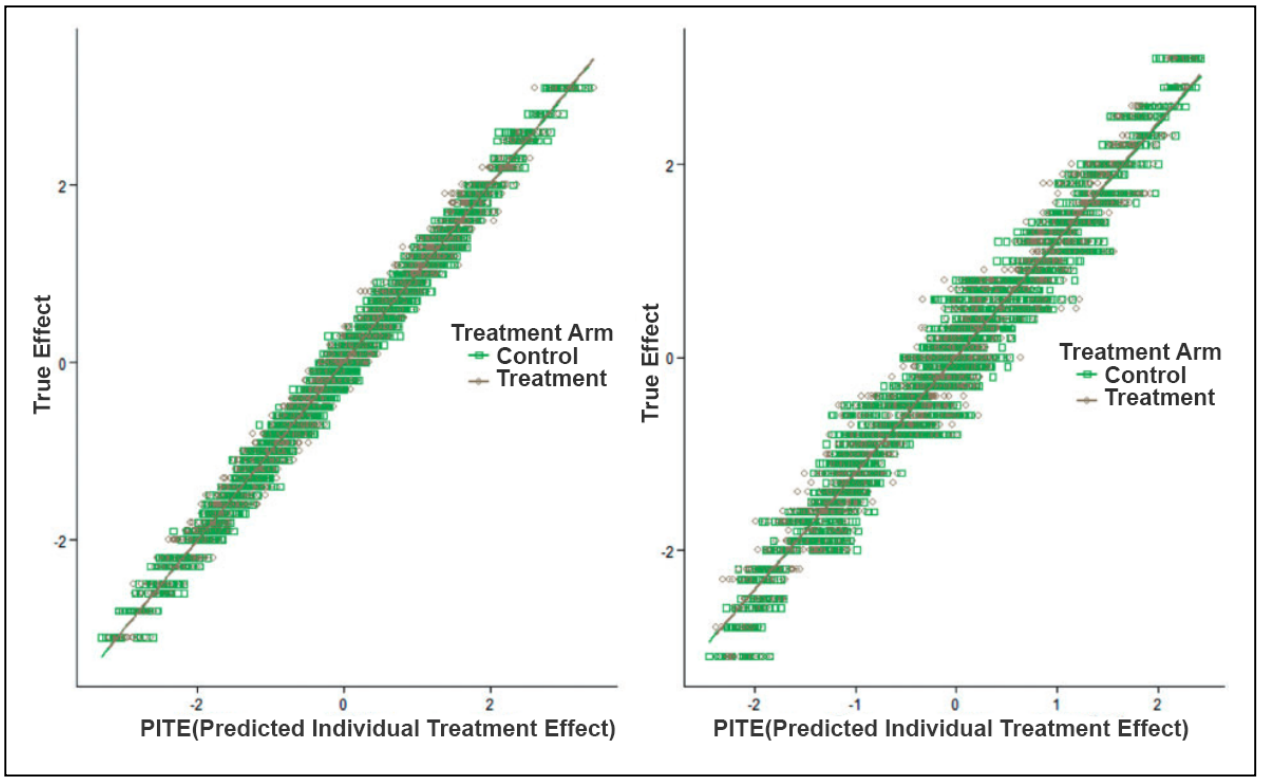
MI is a flexible method for handling missing data that should work well in a randomized trial situation to impute treatment outcomes of missing treatments. We use MI with the outcome set to missing for the treatment(s) that the individual was not assigned and impute $m > 1$ plausible values based on a large set of observed baseline covariates. The ITE is defined as the average difference between the values of outcome for treatment A and the values of outcome for treatment B across imputations, for which data are now available for every individual. The fact that more than 1 imputation is used preserves the underlying uncertainty in the data. The method we used for MI is known as the chained-equations algorithm.

The simulation in the Lamont et al manuscript. In Lamont et al,[2] we created a rather simple data generation model comparing an active treatment with a control treatment. The outcome in the control treatment was purely random, with the variability of the outcome equaling 1. We then generated 7 binary variables; half of the individuals would be zero and half would be 1 on each of these variables. We then generated their true TE under the active treatment using the following formula:

$$TE = -1.3X_1 - 1.2X_2 - .6X_3 + .3X_4 + .5X_5 + 1.1X_6 + 1.2X_7$$

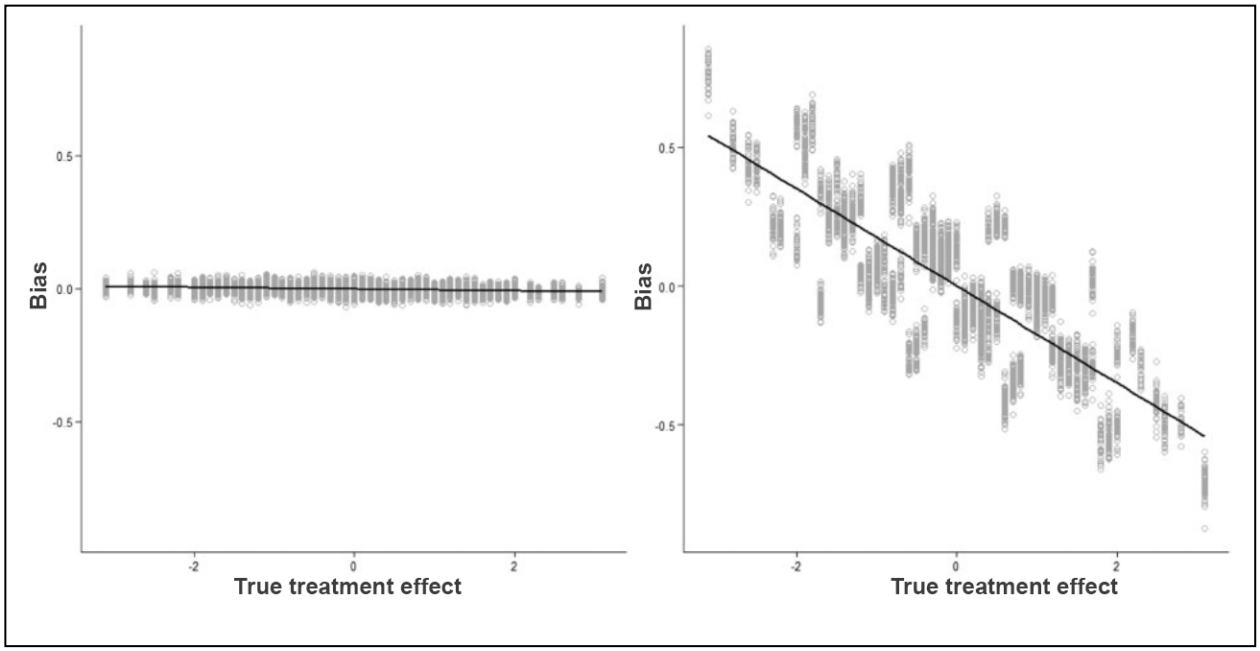
Each individual's observed outcome under the active treatment was the true TE plus a random term with variance equal to 1. The results of the simulation show that both methods worked to estimate the true TE; however, there appeared to be more variability in the RF-based estimates (see Figure 3). In addition, there is a tendency at the tails of the true TE for the RF predictions to be biased toward ~~zero~~0. This bias can be seen clearly in Figure 4.

Figure 3. Relationship ~~b~~Between ~~t~~True TE and ~~p~~Predicted TE[a]



[a]The left-hand plot shows the ~~multiple imputation~~MI approach, and the right-hand plot shows the ~~random forest~~RF approach.

Figure 4. Bias in the 2 Mmethods (MI and RF)[a]



Abbreviations: MI, multiple imputation; RF, random forest.
[a]The left-hand plot shows the MI approach, and the right-hand plot shows the RF approach.

Summary. This simulation showed that both methods, the MI approach and the RF approach, can be used to estimate ITEs using clinical trial data. The predictions were less variable and less biased using MI. It should be noted that the RF is unbiased across the sample but appears to be more biased toward zero in the tails of the true TE. This is consistent with the known edge bias in nearest neighbor (and, in particular, RF) estimators.[6] In this simulation, it appears that MI is superior to RF as an estimator of ITEs. However, it should be kept in mind that the model used to generate the data met all the assumptions for MI to work well, which may not be true for real clinical data. In addition, MI, while a useful method when there are very few predictors, will not do well when there are numerous predictors, such as in the case when genetic data are included in the prediction. Our study 1 expands the data-generating models to be more complex and looks at additional methods for generating ITEs. We believe under different circumstances that different methods will be best to use to estimate ITEs.

Methods for Study 1: How Well Does RF Work to Predict How Individuals Will Do in Treatment With More Complicated Forms of Heterogeneity?

Study 1 expands the data generation models used to test our approaches to ITE estimation. This study examines 3 methods of estimating ITEs: MI, RFs, and Bayesian additive regression trees (BART).[3] Like RFs, BART is tree based and has tuning parameters. It is a Bayesian procedure because users provide a guess as to the correct tuning values, but the algorithm combines this guess with what the data show to be the correct tuning value. The data generation algorithms are as follows:

D1: only random error and heterogeneity

$$F(X) = -0.65 * X1 - 0.6 * X2 - 0.3 * X3 + 0.15 * X4 + 0.25 * X5 + 0.55 * X6 + 0.6 * X7 + 0 * X8$$

D2: heterogeneity with square and 2-way interactions

$$F(X) = -0.47 * X21 - 0.6 * X2 - 0.31 * X3 * X9 + 0.15 * X4 + 0.25 * X5 * X8 + 0.55 * X6 - 0.43 * X27$$

D3: heterogeneity with cubic and 3-way interactions

$$F(X) = -0.17 * X31 - 0.6 * X2 - 0.30 * X3 * X9 * X10 + 0.15 * X4 + 0.25 * X5 * X11 * X12 + 0.55 * X6 + 0.16 * X37$$

D4: heterogeneity with square and nonlinear interactions

$$F(X) = -0.873 * X1 * XL1 - 0.891 * X2 * XL2 + 0.30 * X3 + 0.15 * X4 + 0.25 * X5 + 0.55 * X6 + 0.772 * X7 * X8 * XL1$$

D5: heterogeneity with cubic and nonlinear interactions

$$F(X) = -0.873 * X1 * XL1 - 0.891 * X2 * XL2 - 0.305 * X3 * X9 + 0.153 * X4 * X10 * X11 + 0.253 * X5 * X12 + 0.564 * X6 * X15 * X16 + 0.772 * X7 * X8 * XL1$$

where $X_1, \ldots, X_{16}$ were drawn from a multivariate normal standard $N(0, 1)$ and covariates X_{L1} and X_{L2} were generated as indicator variables for an observed variable (ie, a variable in the feature set) being inside or outside, respectively, of the 33% and 66% percentile cutoffs.

The simulated data sets included 5000 participants, and there were 1000 data sets simulated. We compared the distributions of bias and root MSE (RMSE) of the estimates across the 1000 simulated data sets.

Methods for Study 2: What is the Impact of the Magnitude of Heterogeneity on the Performance of These Estimators?

The purpose of study 2 was to explore how these ITE estimates performed as the level of heterogeneity and the sample size changed. In addition, we examined the impact of prognostic covariates that are predictive of outcome across all treatment variables. As part of this study, we explore the use of various effect size measures for HTE.

Our general model includes 2 types of predictors of outcomes. If we have prognostic covariates, *Z*, which affect *Y* in the same manner regardless of treatment, and predictive covariates, *X*, which differentially affect *Y* depending on treatment assignment, then the potential outcomes for each individual can be represented as:

$$Y_{i0} = E\left(Y_i \mid X_i, Z_i, T_i = 0\right) + \varepsilon_{i0}$$

$$Y_{i1} = E\left(Y_i \mid X_i, Z_i, T_i = 1\right) + \varepsilon_{i1}.$$

Note that either *X* or *Z* (but not both) may include a constant term. Given that this is a randomized study, it seems reasonable to assume that ε_{i0} and ε_{i1} are drawn from the same distribution, in which case:

$$Y_{i0} = E\left(Y_i \mid X_i, Z_i, T_i = 0\right) + \varepsilon_i$$

$$Y_{i1} = E\left(Y_i \mid X_i, Z_i, T_i = 1\right) + \varepsilon_i.$$

If we further assume that *X*, *Z*, and *T* are independent of each other, $Z_i \perp X_i \perp T_i$, and assume linearity of the conditional expectations, then:

$$Y_{i0} = E\left(Y_i \mid Z_i\right) + \varepsilon_i$$
$$Y_{i1} = Y_{i0} + E\left(Y_i \mid X_i\right),$$

and we could describe *Y* as conditional on *T* as:

$$Y_i = E\left(Y_i \mid Z_i\right) + E\left(Y_i \mid X_i, T_i\right) + \varepsilon_i$$
$$Var\left(Y_i\right) = Var\left(E\left(Y_i \mid Z_i\right)\right) + Var\left(E\left(Y_i \mid X_i, T_i\right)\right) + Var\left(\varepsilon_i\right)$$
$$Var\left(Y_i\right) = \sigma_z^2 + \sigma_x^2 + \sigma_\varepsilon^2$$

where $\sigma_x^2 =$ **variance associated with treatment heterogeneity**
$\sigma_z^2 =$ **variance associated with prognostic covariates**
$\sigma_\varepsilon^2 =$ **residual variance (irreducible noise)**

Then, the ITE is:

$$\tau\left(X_i\right) = \left(Y_{i1} - Y_{i0}\right).$$

Given our framework, we can describe the variability in *Y* according to whether it is associated with *X*, *Z*, or residual error.

Definition of effect size. Standard definitions of effect size tend to focus on mean differences across treatments with a standardization using some measure of variance, normally the standard deviation. These effect size measures can be transformed into a proportion of variance explained by examining how much of the variability in outcome is explained by the mean differences. Because HTE implies variability in outcomes, an effect size measure for HTE logically would focus on ratios of variances. One measure of effect size might be the simple R^2 for the amount of variance that is explained by observed characteristics:

$$R^2 = \frac{\sigma_x^2 + \sigma_z^2}{\sigma_x^2 + \sigma_z^2 + \sigma_\varepsilon^2}.$$

However, this confounds the heterogeneity and prognostic predictors. Another possible measure is the simple proportion of the variance associated with heterogeneity:

$$R^2(x) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_z^2 + \sigma_\varepsilon^2}.$$

Of course, in simple multiple regression analysis, prognostic control variables actually increase statistical power and therefore are removed from the denominator:

$$\frac{R^2(x)}{1 - R^2(z)} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\varepsilon^2}.$$

Finally, we could also look at the ratio of the variance of heterogeneity to the residual variance:

$$h^2 = \frac{\sigma_x^2}{\sigma_\varepsilon^2}.$$

Simulation design. The purpose of this simulation is to understand how the magnitude of heterogeneity affects the properties of the predicted ITEs (PITEs). Simulation models with different types of HTE were used to assess the performance of the different estimation methods. Each specification of the simulation was repeated with 1000 data sets generated for each specification. Baseline covariates and the true TE were set to be the same across the 1000 repetitions; however, random error was varied. This established a scenario where the same individual was repeated, allowing for intuitive interpretations about the number of times an individual's predicted value reflected the true TE. There were several repetitions of the simulation with different sample sizes and variance settings. Data were generated in R software, version 3.3.3 (R Foundation for Statistical Computing).

The primary assumption here is that the independent variables are uncorrelated and independently drawn for both continuous and binary types of variables, as follows:

$$X_1, X_2, X_3 \ldots X_8 \sim N(0,.25)$$
$$Z_1, Z_2, Z_3 \ldots Z_6 \sim N(0,.25)$$
$$X_{21}, X_{22}, X_{23} \ldots X_{29} \sim Bernoulli(.5)$$
$$Z_{21}, Z_{22}, Z_{23} \ldots Z_{27} \sim Bernoulli(.5)$$

Note that the variance of continuous measures was chosen to equalize variability across the continuous and dichotomous/binary predictor variables. Distribution of the features is a design factor in the simulations; simulations are done wherein all *Z* and *X* are either continuous variables or all binary variables. We decided to compare continuous and binary predictors to compare the impact of edge bias with the 2 types of predictors. We hypothesized that evidence of edge bias would be strongest with continuous predictors because the edges would be less dense, providing fewer neighbors.

We have ~~a number of~~many additional design factors for the simulations, which are all crossed with the type of predictor. That is, the following conditions are repeated twice, once with binary predictors and once with continuous predictors. Due to the choice of variances described previously, the measures of effect sizes within a specification will be identical for continuous and binomial feature sets.

(1) Four levels of HTE:

$$\begin{aligned}
None: &\quad Y_{i1} = Y_{i0} \\
Low: &\quad Y_{i1} = Y_{i0} + .1X_{i1}T_i + .1X_{i2}T_i - .2X_{i3}T_i + .3X_{i4}T_i - .4X_{i5}T_i + .1X_{i6}T_i \\
Medium: &\quad Y_{i1} = Y_{i0} + .2X_{i1}T_i + .15X_{i21}T_i - .3X_{i3}T_i + .4X_{i4}T_i - .6X_{i5}T_i + .2X_{i6}T_i \\
High: &\quad Y_{i1} = Y_{i0} + .3X_{i1}T_i + .2X_{i2}T_i - .4X_{i3}T_i + .4X_{i4}T_i - .8X_{i5}T_i + .3X_{i6}T_i
\end{aligned}$$

(2) Two levels of prognostic predictors (ie, predict *Y* regardless of treatment):

$$Low(none): \quad Y_{i0} = 0Z_{i11} - 0Z_{i12} - 0Z_{i13} + 0Z_{i14} - 0Z_{i15} + 0Z_{i16} + \varepsilon_i$$
$$High: \quad Y_{i0} = .32Z_{i11} - .24Z_{i12} - .4Z_{i13} + .3Z_{i14} - .39Z_{i15} + .41Z_{i16} + \varepsilon_i$$

(3) Three levels of random error:

$$Low: \quad \varepsilon_i \sim N(0,.1)$$
$$Medium: \quad \varepsilon_i \sim N(0,.2)$$
$$High: \quad \varepsilon_i \sim N(0,.6)$$

(4) Three levels of nuisance (nonpredictive) covariates: k = 8, k = 16, and k = 32

(5) Three sample sizes: n = 500, n = 2500, and n = 5000

(6) Four different types of estimators: VT, RF, synthetic RF, and generalized linear model (GLM)

Methods of estimation. The VT method was proposed by Foster et al[1] to estimate outcomes under the potential outcomes framework. In this method, a single RF is trained to regress the outcome Y_i against covariates $\boldsymbol{X}_i$ and treatment condition T_i. The counterfactual outcome estimate for individual i is then obtained by setting up the VT data, that is, the original observed treatment variable $\mathrm{T_i}$ and its counterfactual treatment $1 - T_i$. These 2 counterfactual conditions are used to obtain the 2 outcome estimates $\hat{Y}_i(1)\ and\ \hat{Y}_i(0)$. Finally, the ITE estimate for individual *i* is defined as $\hat{Y}_i(1) - \hat{Y}_i(0)$. Similarly, for the entire data set, $\hat{Y}_{VT}(\mathbf{x}, T)$ indicates the predicted value for data set $(\mathbf{x}, T)$ from the VT forest; the ITE estimate is then defined as

$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}_{VT}(\mathbf{x}, 1) - \hat{Y}_{VT}(\mathbf{x}, 0).$$

Foster et al[1] mention certain tuning parameters, such as including treatment interactions in the design matrix. This is done by training an RF-R, where Y_i is regressed against covariates X_i, treatment condition T_i, and their pairwise interactions with treatment, $\mathbf{X}_i T_i$.

Our CF approach is similar to VT, but rather than estimating a single RF across both treatments, a separate RF is estimated for each treatment including only the cases assigned the

particular treatment. This procedure can be done in 2 different ways. In the first, the standard RF approach is used; in the second approach, synthetic RFs are used.

Our final estimator simply uses a GLM to fit an outcome model for each treatment where only people in the treatment are included in the GLM. These models are then used to generate predicted values for outcome under each treatment for each individual, and the individual treatment prediction is formed as the difference in these 2 estimates.

Methods for Study 3: Can We Estimate Individual Outcomes (and Therefore Create Predicted ITEs) Using Observational, Potentially Confounded Data?

We have extended these approaches to work with observational, potentially confounded data.[5] Because our work here is based on observational data, different individuals have a different likelihood of getting different treatments. This implies that differences in the characteristics of individuals are related to their likelihood of getting a particular treatment, which causes most simple procedures used to estimate TE to be biased due to this confounding. One method to address confounding is to estimate the likelihood of getting each treatment and to create a propensity score, which is the probability of a particular individual getting a particular treatment. One method to estimate average TEs reweights the estimates by the inverse of the propensity score, which equalizes the distribution of the predictive characteristics across treatments.

In this study, numerous different RF approaches were examined for creation of the ITEs. These approaches did not correct the weighting of individuals but did include all the potentially confounding variables in the set of features used to predict each individual's outcome under different treatments. For each sample, a propensity score was also estimated and was used to assess bias and variability. For each procedure, bias and variability (RMSE) were examined, stratified by the propensity score. Therefore, the ITE estimates did not incorporate the propensity score; however, the assessment of bias did stratify on the propensity score to examine if the procedure performed well across the entire distribution of the propensity.

We examined 7 different methods for generating the PITEs. The first method is VT.[1] In this approach, RFs are used to predict outcome based on individual characteristics and a treatment indicator. In a 2-treatment clinical trial, 1 forest created would be created; the treatment indicator would be set to 0 to obtain the prediction for 1 treatment and then set to 1 to obtain the prediction for the other treatment. The second method is a variant of VT, VT interaction (VT-I). In VT-I, rather than just putting the individual characteristics into the procedure, the individual's characteristics are put in with an interaction. In essence, the vector (list) of characteristics, X, is put in twice. It is multiplied ~~one~~ 1 time by T (the treatment indicator which equals 0 or 1), and the other time by (1 – T). The third method is CF. In this approach, separate RFs are estimated for each of the treatments, including only those people who received the individual treatment. The same predictors, X and outcome, Y, are used as in VT. The fourth method is a counterfactual synthetic RF (synCF). This method is like the CF approach but uses synCF.[50] In this method, multiple RFs are estimated using various levels of the tuning parameters (the number of characteristics considered at each branch of the tree and the minimum number of people in the highest or terminal node in the trees). For each of these forests, the outcome is predicted. The final prediction of outcome is a forest that includes all the original characteristics and these new, predicted outcomes (synthetic characteristics). The fifth method is a bivariate imputation approach (bivariate). This approach uses a multivariate unsupervised RF procedure to impute the outcome for the treatment the individual did not receive as a missing data problem.[4] The sixth method is causal RF (causalRF).[6] The causalRF approach builds a forest on half of the data, holding out half for a second stage of the procedure. CausalRF also uses a different rule for choosing cut points. It chooses the cut point that maximizes the treatment difference within a node (a branch of the tree). Once the forest is completed, the predictions are based on the held-out data, using the forest grown on the training data. The final method we compared is BART.[3] In BART, the multiple trees are averaged as in RF, but there is a Bayesian prior set on tuning parameters. The procedure has been described as modeling the regression surface to estimate the potential outcomes (see Hill[3]) and therefore is similar to the VT approach, with BART replacing RF.

We also tested these methods for generating person-specific TE estimates using 3 different data generation models. Twenty covariates were created; the first 11 were normally distributed, with a mean of 0 and variance of 1, and the next 9 were Bernoulli (0-1), with a probability of 0.5 of being a 1 (and 0.5 of being a 0). To simulate the observational nature of the data, we created a logistic model (the equation is in the logit scale) for the probability of treatment equaling 1 (relative to 0):

$$F(X) = -2 + .028X_1 - .374X_2 - .03X_3 + .118X_4 - .0394X_{11} + .875X_{12} + .9X_{13}.$$

The 3 data generation models were

$$f_1(X,T) = 2.455 - 1_{\{T=0\}} * (.4X_1 + .154X_2 - .152X_{11} - .126X_{12}) - 1_{\{T=1,\, g(X)>0\}}$$
$$f_2(X,T) = 2.455 - 1_{\{T=0\}} * \sin(.4X_1 + .154X_2 - .152X_{11} - .126X_{12}) - 1_{\{T=1,\, g(X)>0\}}$$
$$f_3(X,T) = 2.455 - 1_{\{T=0\}} * \sin(.4X_1 + .154X_2 - .152X_{11} - .126X_{12}) - 1_{\{T=1,\, h(X)>0\}}$$

where $1_{\{\}}$ is an indicator function which is 1 if the condition in the brackets is true and 0 otherwise. Further, the 2 functions are

$$g(X) = .254X_2^2 - .152X_{11} - .4X_{11}^2 - .126X_{12}$$
$$h(X) = .254X_3^2 - .152X_4 - .126X_5 - .4X_5^2 .$$

Note that all 3 data generation models are confounded because characteristics that are related to the probability of treatment are related to treatment outcomes, and there is heterogeneity because there are interactions with treatment. Finally, each of these methods and data generation models was examined for estimating the ITEs.

Methods for Study 4: Can These Methods Be Expanded to Survival Outcomes and to Compare Multiple Treatments?

An article currently under review[60] focuses on ischemic cardiomyopathy and uses observational data to make causal treatment comparisons among 4 distinct treatments: (1) CABG, (2) CABG+SVR, (3) CABG+MVA, and (4) LCTx. This approach proposes a method to ensure

that there is full overlap (eg, that treatment comparisons are only made for treatments for which an individual is eligible) and allows expert knowledge to be included in the procedure. There are important differences regarding the ITE estimates from our other work. In what we have described so far, TEs are simple differences in outcomes between 2 treatments. When the outcome is survival, there are different survival curves (or functions) across each treatment. This means that rankings of treatments may reverse over time. For example, treatment A may improve the chance of survival relative to treatment B in the first year after treatment, but in later years, the chance of survival may be greater under treatment B than under treatment A. The assumptions necessary for this approach to estimating ITEs with survival outcomes are described in the next section.

Treatment effect in survival. The formal notational setup for our extension of individual treatment effective analysis to survival data is as follows. Let $\{(\mathbf{X}_1, Z_1, T_1, \delta_1), \ldots, (\mathbf{X}_n, Z_n, T_n, \delta_n)\}$ denote the data, where $\mathbf{X}_i$ denotes the covariate vector for individual i, (T_i, δ_i) is the observed survival outcome, and Z_i denotes i's assigned treatment group, where Z_i is coded as an integer value from $1, \ldots, M$, where $M > 1$ is the total number of available treatments. The individual's survival outcome is composed of the observed survival time $T_i = \min(T^o, C^o)$ and the censoring variable $\delta_i = \mathbf{1}\{T^o \leq C^o\}$, where T^o is the true (potentially unobserved) event time, assumed to be independent of the true (potentially unobserved) censoring time, C^o. We say i is right-censored at time T_i if $\delta_i = 0$; otherwise, the individual is said to have experienced an event at T_i. Following Rosenbaum and Rubin,[51] we provide a definition for strongly ignorable treatment assignment within the survival setting:

> DEFINITION 2.1. *Let $T^o(j)$ and $T^o(k)$ denote the potential outcomes (event times) under treatments $Z = j$ and $Z = k$, respectively. We say that strongly ignorable treatment assignment (SITA) holds, if for all $j \neq k \in \{1, \ldots, M\}$,*
>
> $Z \perp \{T^o(j), T^o(k)\} \mid \mathbf{X}$.
>
> *In other words, if SITA holds,* $P\{T^o(j) \in \cdot \mid Z, \mathbf{X}\} = P\{T^o(j) \in \cdot \mid \mathbf{X}\}$ for $j = 1, \ldots, M$.

Another key assumption in our development is complete overlap. Let $e_j(\mathbf{x}) = P(Z = j|\mathbf{x})$, the propensity score (for treatment assignment). Complete overlap is said to hold for **x** if $0 < e_j(\mathbf{x}) < 1$ for $j = 1, \ldots, M$. With these 2 definitions in hand, we are now ready to define various useful quantities for assessing treatment effectiveness. We begin by providing a definition for ITE in survival settings. Note that our definition of ITE is a function of both **x** and *t*.

DEFINITION 2.2. *The ITE at time t for covariate* **x** *for treatment j over treatment k is defined as follows*:

$\tau_{j,k}(t, \mathbf{x}) = \psi\{S(t(j)|\mathbf{x}), S(t(k)|\mathbf{x})\}$,

where $\psi(.,.)$ *is a known function and* $S\{t(j)\ \mathbf{x}\} = P\{T^o(j) > t|\mathbf{X} = \mathbf{x}\}$ *is the survival function for the potential outcome* $T^o(j)$ *conditioned on* $\mathbf{X} = \mathbf{x}$.

The assumption of SITA ensures that $\tau_{j,k}(t, \mathbf{x})$ is estimable from the observed data. Under SITA, we have

$S(t(j)|\mathbf{x}) = P\{T^o(j) > t|\mathbf{X} = \mathbf{x}\}$

$= P\{T^o(j) > t|\mathbf{X} = \mathbf{x}, Z = j\}$

$= P\{T^o > t|\mathbf{X} = \mathbf{x}, Z = j\}$

$= S(t|\mathbf{x}, Z = j)$

where $S(\boldsymbol{t} \mid \mathbf{x}, Z = j)$ is the survival function for T^o conditioned on $\mathbf{X} = \mathbf{x}$ and $Z = j$. Thus, under SITA, the survival function equals $S(\boldsymbol{t(j)} \mid \mathbf{x})$, which ensures that the potential outcome survival function is estimable. In general, these 2 values may not be equal without this assumption.

Under SITA, we can now write the ITE as follows:

(2.1) $\tau_{j,k}(t, \mathbf{x}) = \boldsymbol{\psi\{S(t|\mathbf{x}, Z = j), S(t|\mathbf{x}, Z = k)\}}$.

Given an estimator $\hat{S}(t \mathbf{x}, Z)$ for the survival function, it is clear we can estimate the ITE by calculating ψ using the estimated survival function. Examples of $\psi(.,.)$ that can be used to define the ITE include

$$\tau_{j,k}^{(1)}(t,x) = S(t \mid x, Z=j) - S(t \mid x, Z=k), \quad (2.2)$$

where $\psi(a, b) = a - b$, so that $\tau_{j,k}(t, \mathbf{x})$ is the difference of 2 survival curves. Another way to measure ITE is through survival curve domination,

$$\tau_{j,k}^{(2)}(t,x) = 1_{\{S(t|x,Z=j) > S(t|x,Z=k)\}},$$

which corresponds to $\psi(a, b) = \mathbf{1}\{a > b\}$.[6]

Assessing overlap in treatment assignment. To assess overlap in treatment assignment, we used and compared 3 approaches for determining treatment eligibility. In this section, we report the comparability of the resulting groups in terms of the balance of covariates and overlap (when using propensity scores). All 3 of these methods use expert knowledge (E_{ij} is an eligibility indicator for person *i* for treatment *k*) but to differing degrees. Note that in ischemic cardiomyopathy, expert knowledge cannot be considered a gold standard, so although it is a guideline, there is room for improvement on it. The first 2 methods use E_{ij} to calibrate a cut point for exclusion. The third method actually uses a direct prediction of expert knowledge. In the first method (RF-C), an estimated probability of receiving each of the treatments is estimated as a function of baseline features using RFs for classification. In the second method (RF-D), a new distance-based RF procedure *is* used, wherein the estimated probability of RF-C is used as the outcome, and individual characteristics are used as predictors. Once the RF has been estimated, the distance of each person from every other person is estimated, where distance is based on the proportion of the branches of the tree that the 2 individuals traversed in common. These distances are of course indexed by individuals (*i* and *j*), but also by treatment received, *k*. The likelihood of being assigned to treatment *k* for individual *i* is the sum of (1-distance) from person *i* to all persons who receive treatment *k* divided by the sum of (1-distance) of person *i* from all individuals in the sample. For both of the first 2

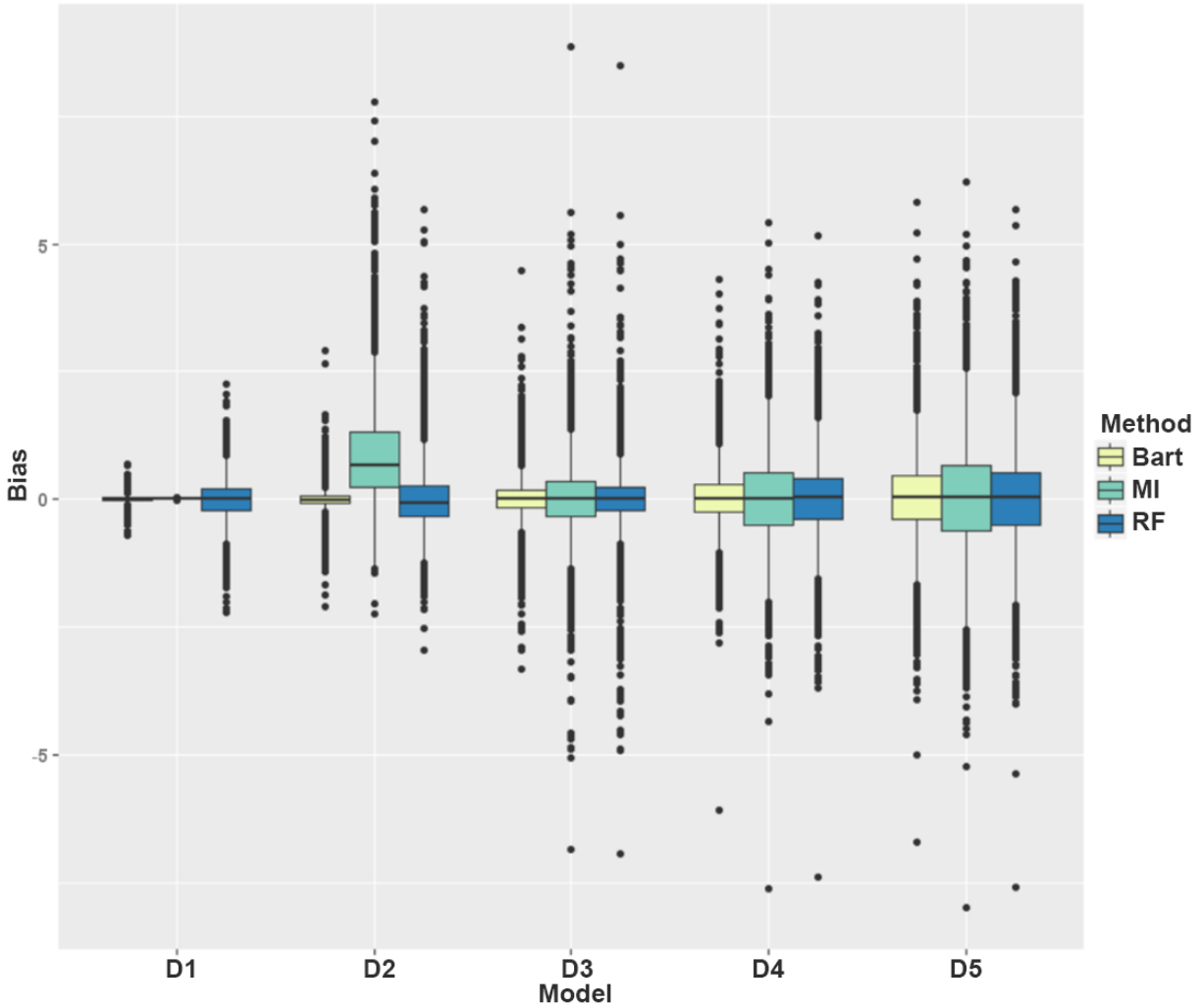
methods, a cutoff score for inclusion in the sample or group to be predicted is calculated by choosing the cutoff score that minimizes the misclassification error using expert knowledge (clinical guidelines from the American College of Cardiology/American Heart Association) as the target (ie, the cutoff is chosen to minimize the occasions in which the decision to include someone disagrees with expert opinion). The final approach uses multivariate RFs (MRFs) according to expert knowledge. Each MRF is calibrated, and a cut point for inclusion in sets of eligibility is found by minimizing the misclassification error (of what the individual actually received).

RESULTS

Study 1

Recall that study 1 extended the findings of Lamont et al[2] by examining multiple more-complex data generation models with higher-order interactions and nonlinear interactions. These simulations showed that in at least 1 simulation (D3), MI did not do as well as RFs in terms of bias. Further, there is less spread in the distribution of bias in RF and less still in BART across the data generation models (Figure 5).

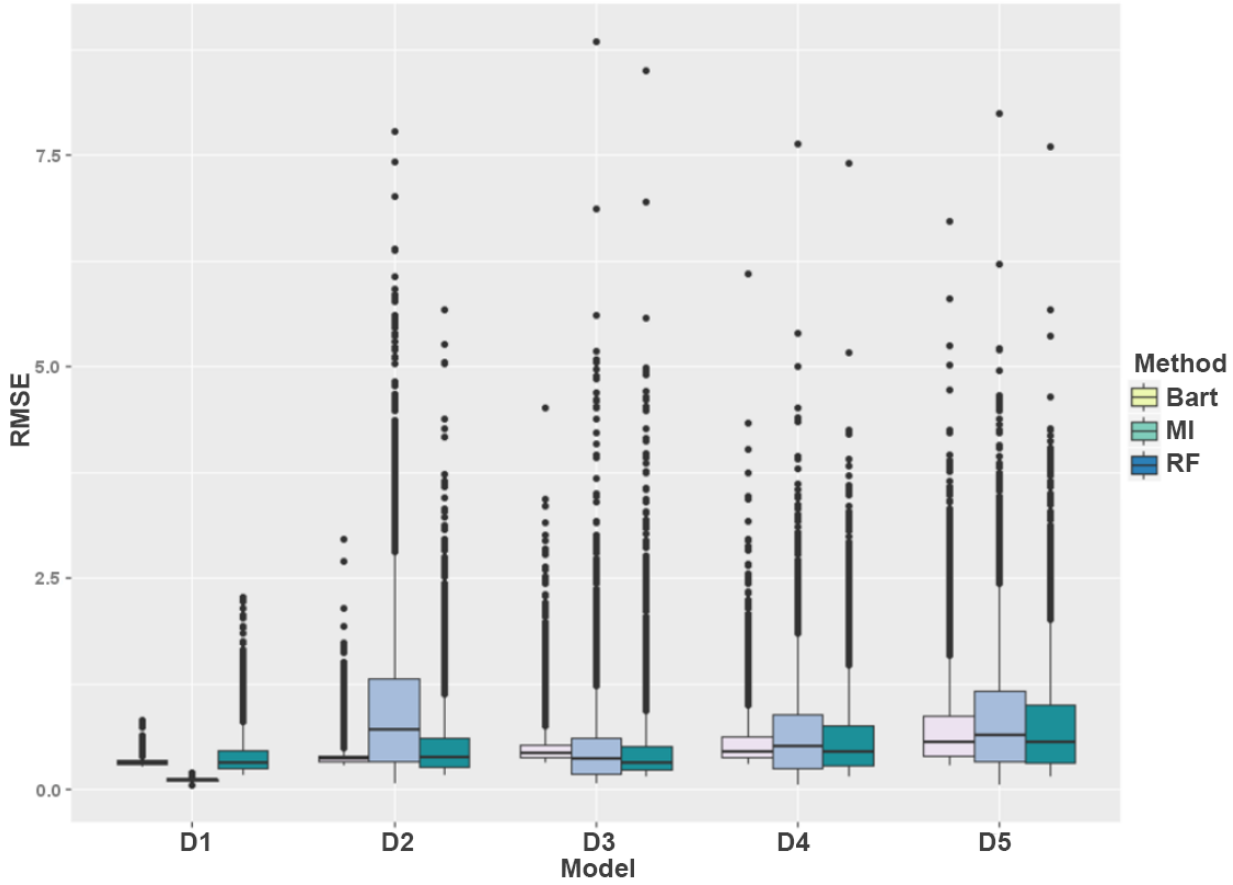
Figure 5. Bias Comparison Across All Simulations With MI, RF, and BART Aapproaches (Distribution of Bias Across Replications in the Ssimulations)



Abbreviations: BART, Bayesian additive regression trees; MI, multiple imputation; RF, random forest.

As shown in Figure 6, the ranking of the distributions of RMSE across simulations seems to hold across data generation models. MI has at least a slightly higher mean RMSE and a wider spread of RMSE across the data generation models than does RF. BART appears to have the lowest spread of RMSE across the data generation models.

Figure 6. RMSE Comparisons Across Simulations



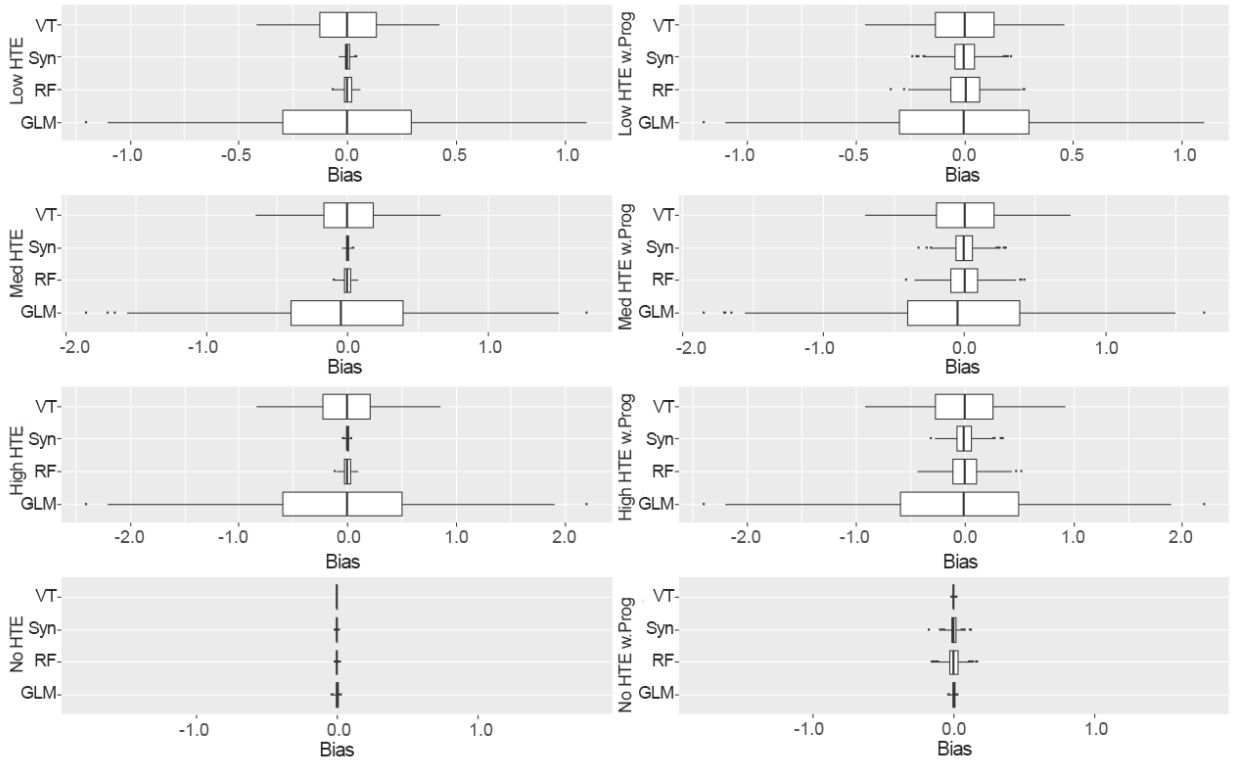
Abbreviations: BART, Bayesian additive regression trees; MI, multiple imputation; RF, random forest; RMSE, root mean square error.

Study 2

Recall that study 2 was a simulation study that varied the amount or size of the HTE, the amount of variability in other prognostic covariates (which predict outcomes across treatment), sample size, and the amount of residual variability. It compared the RF, synthetic RF, and GLM and also varied the distribution of the predictors, either binary or continuous. Figure 7 shows

the distribution of bias across individuals for different levels of heterogeneity and levels of prognostic covariates.

Figure 7. Bias Across Levels of Heterogeneity With and Without Prognostic Covariates[a]



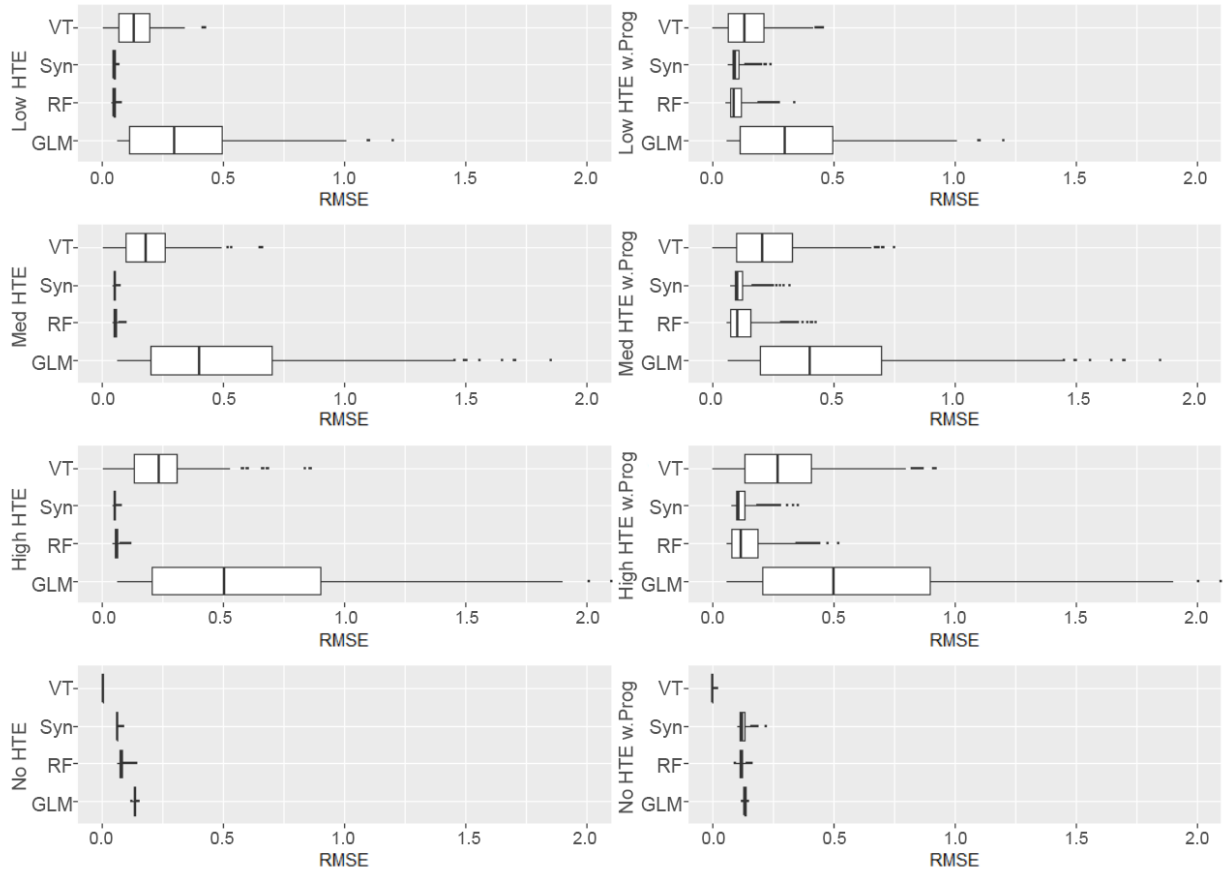
Abbreviations: GLM, generalized linear model; HTE, heterogeneity of treatment effects; Med, medium; w.Prog, with prognostic covariates; RF, random forest; Syn, synthetic RF; VT, virtual twin.
[a]The left panels show the distribution of bias across replications in the simulation at various levels of heterogeneity with no prognostic covariates. The right panels show results at various levels of heterogeneity but with each including the prognostic covariates. Note that the x-axis scale varies by row.

Within each panel, note that the synthetic RF approach has the least amount of bias. Comparing across rows, note that the range of bias increases as heterogeneity increases (ie, the x-axis scale changes across rows). The other striking difference among the panels is that whereas mean bias across the sample stays centered at zero, the spread of bias within the sample increases substantially for the RF estimators when there are prognostic covariates included in the data generation model. Finally, it would appear that GLM estimates have the

largest spread of bias, perhaps due to the lack of model selection (as would occur in boosting, for example).

Figure 8 presents the RMSE for each of the simulations included in Figure 7.

Figure 8. RMSE Across Levels of Heterogeneity With and Without Prognostic Covariates[a]



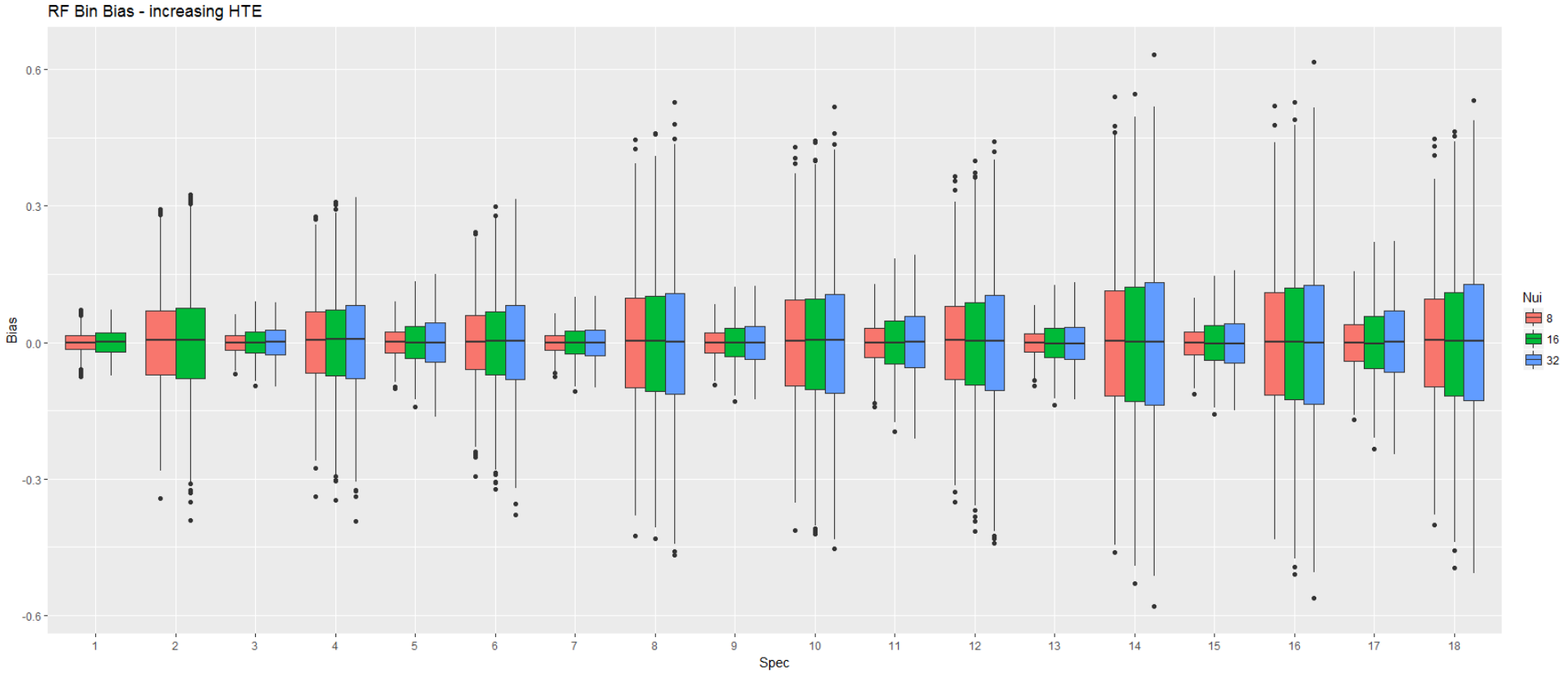
Abbreviations: GLM, generalized linear model; HTE, heterogeneity of treatment effects; Med, medium; w.Prog, with prognostic covariates; RF, random forest; RMSE, root mean square error; Syn, synthetic RF; VT, virtual twin.
[a]The left panels show results at various levels of heterogeneity with no prognostic covariates. The right panels show results at various levels of heterogeneity but with each including the prognostic covariates. Note that the scale of the x-axis varies for the no-HTE row.

From examination within a panel, it is clear that the synthetic RF estimates dominate and have a smaller RMSE than that of the other estimators. There is an increase in the spread of the distribution of the RMSE as both the level of heterogeneity and the variability associated with other prognostic covariates increase. In addition, there is an increase in RMSE (a shift of the distribution to the right) as the variability associated with prognostic covariates increases, though this is less evident in the synthetic RF and RF estimators.

Figure 9 shows all 18 specifications with positive amounts of heterogeneity for the case of binary predictors and using the RF estimation method. Within each specification, there are 3 levels of nuisance variables, 8, 16, and 32.

It is clear that the largest increase in the spread or range of bias across individuals is added when prognostic covariates are included in the data generation model. Increasing the number of nuisance parameters causes a perceptible increase in the spread or range of bias across individuals in all specifications shown. This pattern is also very clear when synthetic RF is used as the estimation strategy (Figure 10). However, the spread or range of bias across individuals is much less in the case of synthetic RF (note the differences in the scale of the y-axis compared with that of Figure 9).

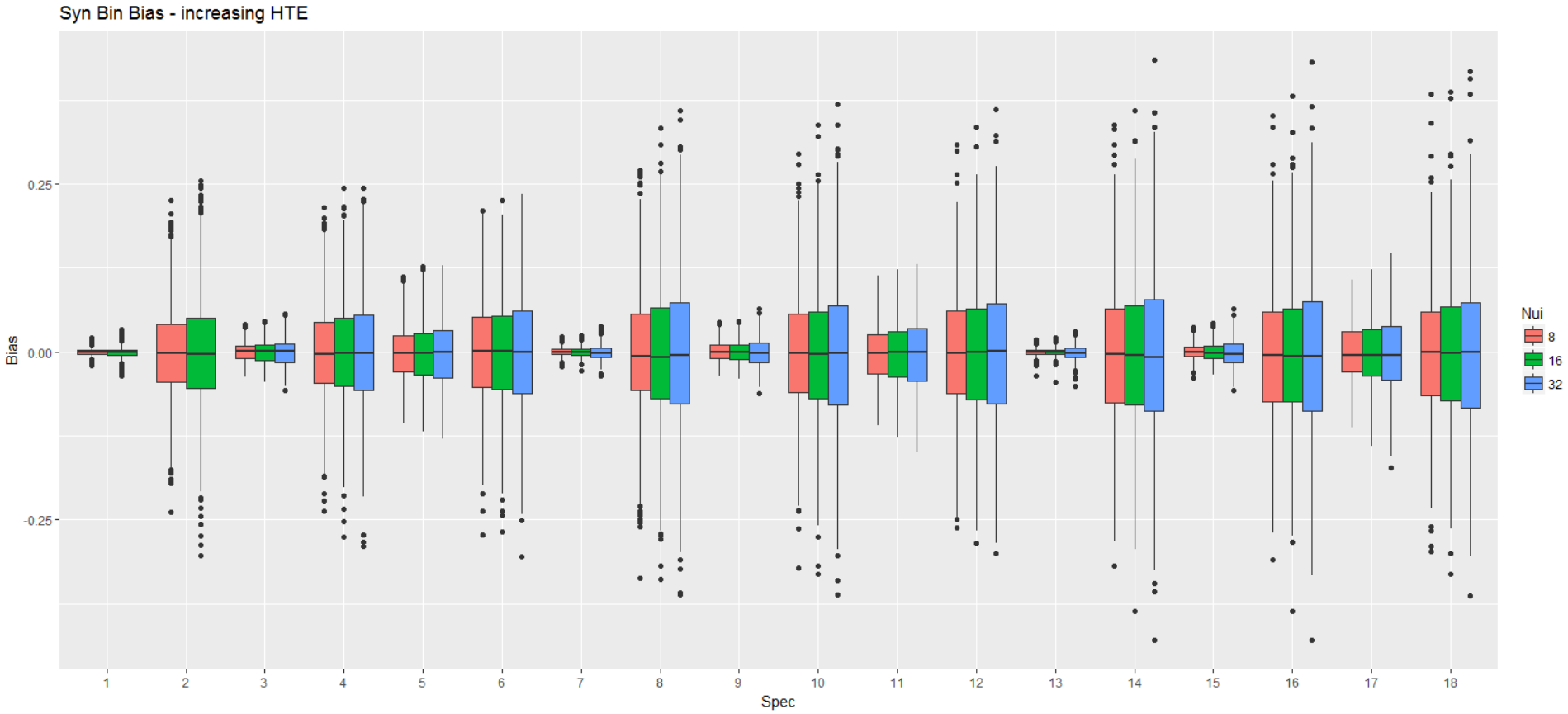
Figure 9. Bias Wwith RF Estimator, All Specifications With Different Numbers of Nuisance Parameters for Binary Predictors



Heter	Low	Low	Low	Low	Low	Low	Med	Med	Med	Med	Med	Med	High	High	High	High	High	High
Res Err	Low	Low	Med	Med	High	High	Low	Low	Med	Med	High	High	Low	Low	Med	Med	High	High
V(Prog)	None	High	None	High	None	High	None	High	None	High	None	High	None	High	None	High	None	High	None	High

Abbreviations: Heter, heterogeneity; HTE, heterogeneity of treatment effects; Med, medium; Nui, nuisance; Res Err, residual error; RF, random forest; V(Prog), variability in prognostic covariates.

Figure 10. Bias With Synthetic RF Estimator, All Specifications With Different Numbers of Nuisance Parameters for Binary Predictors

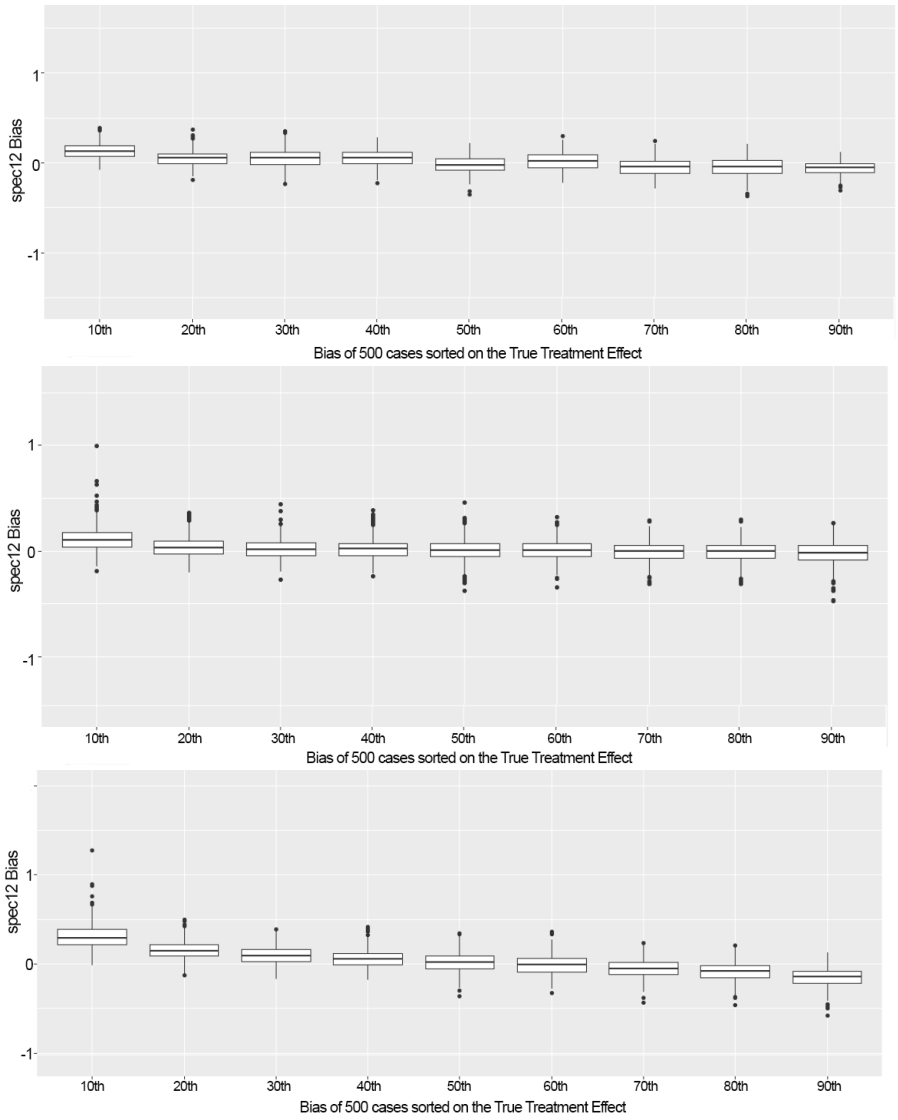


Heter	Low	Low	Low	Low	Low	Low	Med	Med	Med	Med	Med	Med	High	High	High	High	High	High
Res Err	Low	Low	Med	Med	High	High	Low	Low	Med	Med	High	High	Low	Low	Med	Med	High	High
V(Prog)	None	High	None	High	None	High	None	High	None	High	None	High	None	High	None	High	None	High

Abbreviations: Heter, heterogeneity; HTE, heterogeneity of treatment effects; Med, medium; Nui, nuisance; Res Err, residual error; RF, random forest; Syn, synthetic; V(Prog), variability in prognostic covariates.

In Figure 11, we can see that when predictors are continuous, the edge bias associated with RF is most pronounced. Note in the first decile (which has the most negative PITE) that the bias is clearly greater than zero and as you go to the middle of the PITE distribution that bias is near zero. Then, as you move toward the highest decile (where PITEs are most positive), the bias is centered below zero. As we hypothesized, this bias is less apparent in the binary case. We believe this is because with all binary predictors, there is a clumping of the cases in the feature space (at least with finite predictors). Finally, if you compare the bottom panel, which was estimated by synthetic RF, with binary predictors to the same binary predictor specification estimated with RF, there appears to be little difference in the amount of this edge bias.

Figure 11. Bias by Ordered Decile of the True TE



-Abbreviations: RF, random forest; spec12, specification with medium heterogeneity, high residual variance and high variability associated with prognostic covariates; TE, treatment effect.

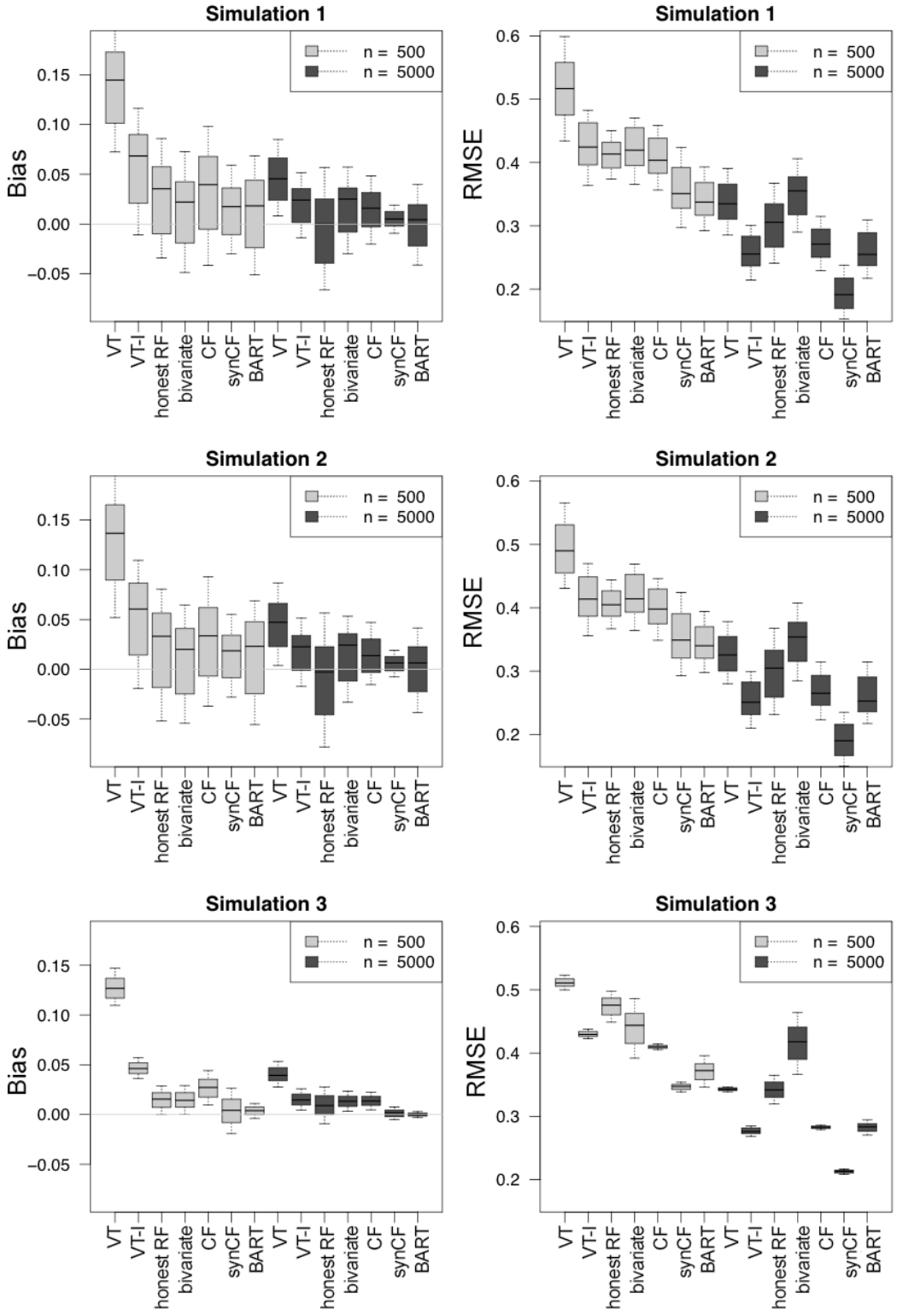
Study 3

Recall that study 2 extended the ITE estimate methodology to work with observational confounded data and compared 7 different methods of forming the ITE estimates.

Figure 12 shows that the synCF and BART approaches have the lowest bias, and although both have low RMSE, synCF has the most precise estimates based on RMSE. These simulations show that our methods can be used to measure potential confounds and report the

construction of propensity scores. As long as all confounding variables are included in the RF prediction of counterfactual outcomes, the procedure will appropriately account for the confounding (without construction of propensity scores). Creating a propensity score machine would be useful for future implementation of the procedure, however. The propensity score machine can be used to ensure that future patient characteristics result in a propensity score that is neither 0 nor 1 (ie, not assigned to 1 treatment with certainty). This study has been published in the *Journal of Computational and Graphical Statistics*.[5]

Figure 12. Bias and RMSE for Ssimulations of Confounded Observational Data



Abbreviations: BART, Bayesian additive regression trees; CF, counterfactual random forest; RF, random forest; RMSE, root mean square error; synCF, counterfactual synthetic random forest; VT, virtual twin; VT-I, virtual twin interaction.

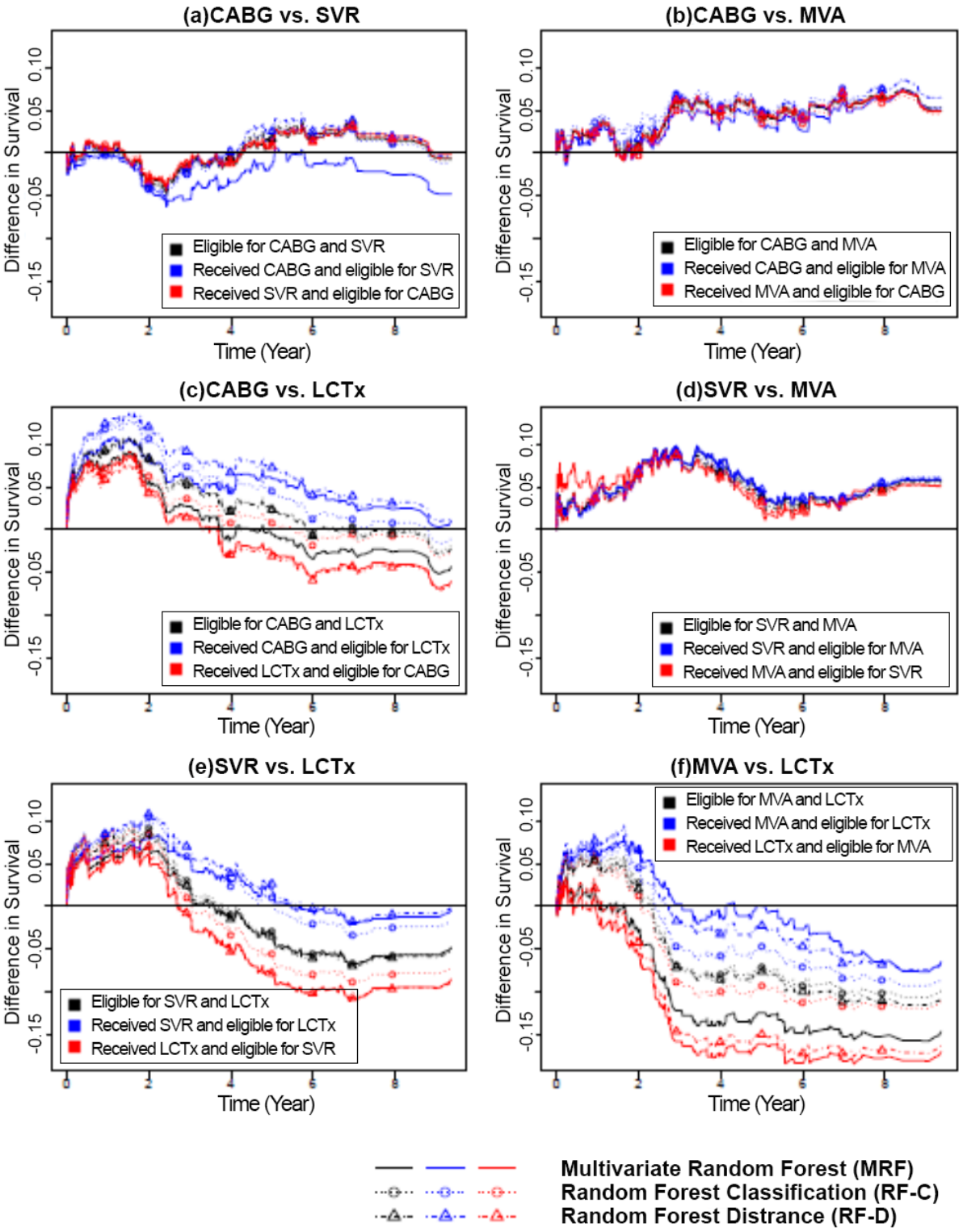
Study 4

Recall that study 4 extended the methods to create estimates of ITEs to survival outcomes. We also described the following 3 methods for determining whether an individual was suitable to have their treatment predicted for a particular treatment: RF-C, RF classification with calibration to expert knowledge; RF-D, RF distance with calibration to expert knowledge; and MRF, multivariate RF with direct estimation of expert knowledge. In a comparison of these 3 approaches in terms of misclassification error, RF-C and RF-D were very close (0.32 and 0.35, respectively, across all treatments), and MRF had the lowest misclassification error (0.13). Given that expert knowledge cannot be considered a gold standard, all 3 of these approaches were examined in subsequent analyses.

There is not a single number but rather a function comparing survival across each pair of treatments compared. To summarize the data, we compare these functions across averages across the sample. The ITEs can be used to create average TE (ATE) estimates as well as ATE in the treated (ATT) estimates. These are also both functions, rather than single numbers. Figure 13 shows the results of pairwise comparisons based on the ATE and ATT. When the lines cross the black lines at zero, the ranking of the 2 treatments reverses at that time. From looking at panel (d), SVR appears everywhere to be superior to MVA, but this is really the only panel for which this is the case. Other things to note about the figure include that the sample used to compare the 2 procedures depends on and varies by the procedure used to determine eligibility for the treatment.

Because the survival estimate for an individual is a full survival curve across time, it is more difficult to show the individual treatment estimates in a summary fashion. We can easily show these estimates at a particular time of follow-up. Figure 14 shows individual estimates for 5 years after treatment, with confidence intervals calculated using subsampling.[61]

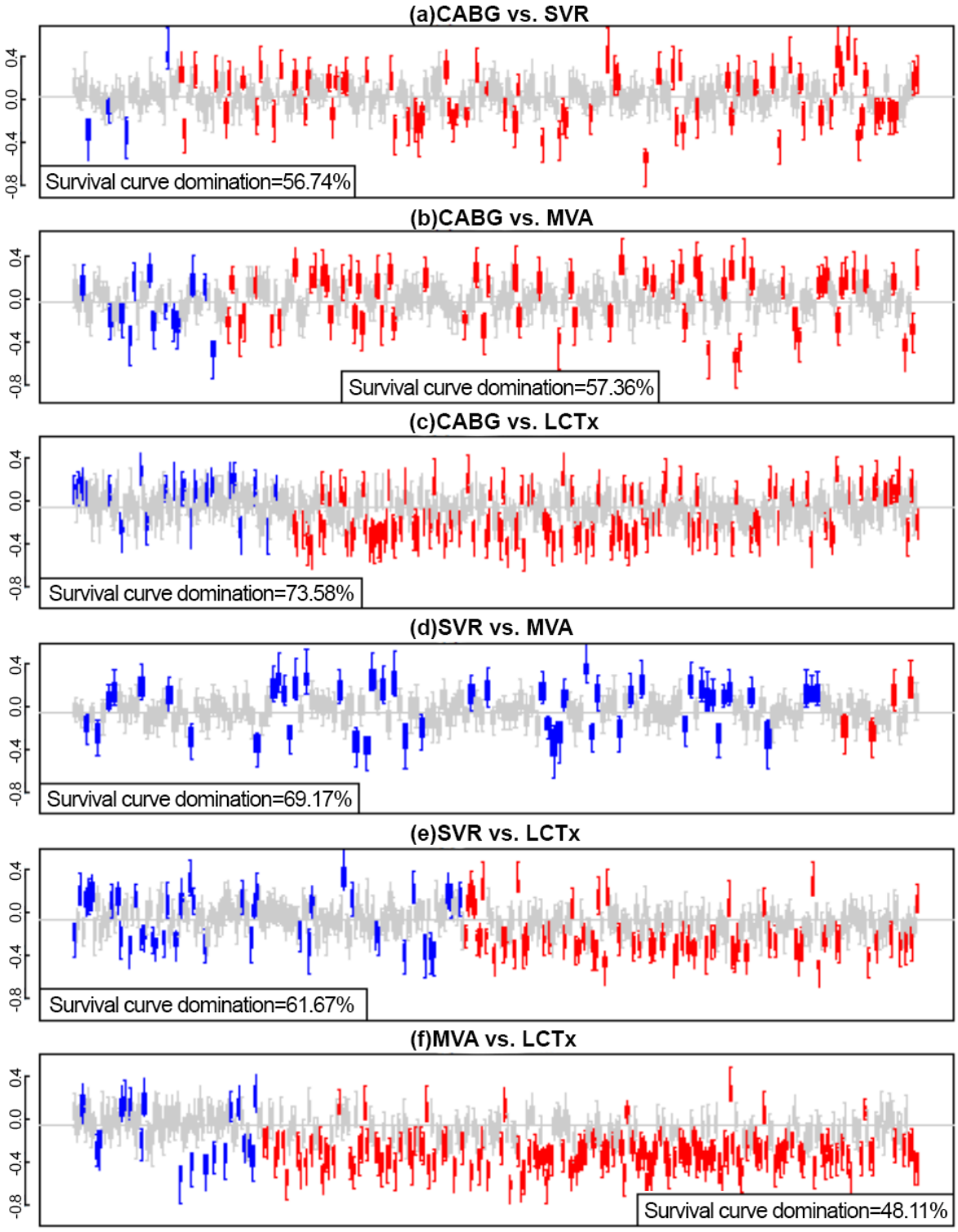
Figure 13. Pairwise Ccomparisons of Treatment Options on ATE and ATT[a]



Abbreviations: ATE, average treatment effect; ATT, ATE in treated; CABG, coronary artery bypass graft; LCTx, listing for cardiac transplantation; MVA, mitral valve annuloplasty; SVR, surgical ventricular restoration.

[a]ATE and ATT eligibility was determined by 3 methods: RF-C, RF-D, and MRF. Black lines are ATE, blue lines are ATT for the first treatment listed in a panel, and red lines are ATT for those in the second treatment listed.

Figure 14. Confidence Iintervals for ITEs at Ttreatment Time= 5 Yyears[a]



Abbreviations: CABG, coronary artery bypass graft; ITEs, individual treatment effects; LCTx, listing for cardiac transplantation; MVA, mitral valve annuloplasty; SVR, surgical ventricular restoration.

[a]Blue indicates significant TE ($P < 0.05$) for the treatment mentioned first in each panel. Red indicates significant TE ($P < 0.05$) for the second treatment mentioned in each panel. Patients are randomly ordered within blue and red.

Results of Aim 2

Aim 2 was to develop user-friendly software integrated into randomForestSRC. There have been multiple releases of the randomForestSRC program incorporating our work on this project. A summary of the improvements to randomForestSRC includes the following:

1. Enhanced synthetic forests (function rfsrcSyn()). This corrected issues with colnames of test set synthetic features. OOB data are now incorporated into the forest grow process, leading to overall better performance.

2. Introduced bootstrap="by.user". This allows the user to control the bootstrap process and maintain in-bag/OOB tree membership across versions of synthetic forests across variation in tuning parameters.

3. Introduced samp, sampsize, and samptype options. These are used for class-imbalanced data (where 1 class is far away from 50% of the sample). The Project AWARE data are class imbalanced when incidences of STIs are the outcome, for example, because incidences are around 10% to 2%, which is relatively imbalanced (far from 50%).

4. Developed subsampling methodology, with function subsample (). This is used for estimates of standard errors and can create standard errors for treatment effectiveness. Another application is creating standard errors and confidence intervals for variable importance.

5. Introduced conditional quantiles for a regression forest. This applies to both univariate and multivariate forests and can be used in both training and testing. The function returns the conditional quantiles for the target outcome. It is used when the treatment is continuous.

6. Added a configure file to source package to allow more accessible OpenMP parallel execution on systems that support it.

7. Added staggered interaction data (SID) clustering function. SID clustering is used for semi-supervised analyses to uncover HTE groups. This is an outgrowth of the split-merge algorithm and the unsupervised RF proposed in this project. SID clustering creates an enhanced SID feature space by sidification of the original variables. Sidification translates the variables to have nonoverlapping domains. Sidification results in (1) SID main features, which are the original features that have been shifted in order to make them strictly positive and staggered so all of their ranges are mutually exclusive; and (2) SID interaction features, which are the multiplicative interactions formed between every

pair of SID main features. MRFs are then trained to predict the main SID features using the interaction SID features as predictors.

DISCUSSION

Our research has shown that RF can be successfully used to estimate ITEs, and in cases where heterogeneity is generated using a linear model, these estimates perform as well as parametric approaches such as GLM (study 2). This extends the work of Foster et al,[1] who showed that the RF approach worked in the VT approach but did not compare across algorithms. We also saw that the synthetic RF approach outperformed GLM with linear data generation models, and to our knowledge, no one has examined the use of synthetic RF for the estimation of ITEs. We believe that the improved performance of synthetic RF over that of RF may be caused by smoothing the predictions by including information from multiple forests, each with different tree properties. Both RF and synthetic RF are nonparametric approaches that should do well with any distribution of data and with most models generating heterogeneity. Given the superior performance of the synthetic RF approach, we would recommend its use over the standard RF approach.

We also saw in study 1 that RF outperformed the MI approaches when there were higher-order interactions and nonlinear interactions in the generating mechanism for HTE. This is consistent with our belief that different estimation methods of ITEs may be required for different applications of this approach. The nonparametric basis for RF (and synthetic RF) implies that it should perform well across many different data-generating mechanisms; however, there are undoubtedly particular heterogeneity-generating mechanisms for which other procedures may have better performance than RF-based estimates. We greatly expanded the types of algorithms that we examined beyond the simple RF approach because of our belief that different algorithms will be useful for our approach under different circumstances. Nevertheless, there remain numerous methods that we have not compared (eg, neural nets, support vector machines, boosted decision trees).[62] A challenge for future research is to determine and understand the data conditions under which a specific tool should be used. This is, of course, complicated in real data because the data-generating mechanism is unknown. It may be that an ensemble approach, in which estimates across methods, including methods

which themselves are an ensemble, such as RF, would work well,[63] and future research should explore this.

Our research into the impact of the level or amount of heterogeneity (study 2) showed that as the level of heterogeneity increased, the amount of bias and the RMSE of the individual estimates increased, holding the sample size constant. With consideration, this seems logical. The identification of ITEs exploits individual characteristics or features to make the prediction; our predictions are conditional on the individual's features. With sample size fixed and the distribution of features fixed (as was done in our simulations), increasing heterogeneity implies that within each region of the feature space, there will be more variability in the outcome; hence, our estimates of the "individual" TE will have more variability. If the mean of this region is assigned for all of those individuals that fall within the region, then the increased heterogeneity within the feature region means that the average bias in the region (ie, the deviation of the ITE from the individual's true TE) will increase, as will the RMSE.

The other interesting phenomenon observed in study 2 was that as we added in prognostic variables (recall that prognostic variables have an identical effect on outcomes across all possible treatments), the levels of both bias and RMSE increased, holding all other factors constant. We believe that this is caused by the error in predicting the prognostic impact. This is different from how we think of control variables in a regression framework, however. Controlling for variables in regression reduces residual variance and sharpens our inference on remaining variables or features (ie, it increases statistical power and helps us in our prediction). Here, prognostic factors that predict outcomes across treatments do not appear to help us in the prediction of ITEs and, in fact, hinder our efforts. This may be caused by the added noise of estimating the impact of these prognostic factors in different forests (1 for each treatment to be compared). Methods that focus on treatment interactions, such as those of Su et al,[64] may not show this same pattern.

It is also important to point out that the GLM-based predictions in these simulations had more bias across the individual predictions and larger RMSE than did the RF-based estimates, despite our data generation models having a linear form. This is likely due to the numbers of

nonpredictive covariates in the simulations. Although these simulations clearly had fewer nuisance variables included than would be found in a genetic study, for example, there were enough that the noise of estimating these extra (noninformative) coefficients degraded the performance of the GLM estimates. This could be addressed by some form of regularization procedure. Indeed, numerous investigators have examined the use of regularized models (for model selection) as a part of a procedure for either ITE prediction or subgroup discovery.[26,65-68]

This research has also shown (study 3) that the RF approaches to predicting ITEs perform well using observational data if all confounding variables are observed and included in the feature set. Interestingly, this did not require inclusion of the propensity score in the feature set or stratification or matching on the propensity score. There are times, particularly with misspecification or measurement error, that propensity scores and disease risk scores can outperform regression adjustments of confounds.[69,70] Research on variable selection in estimation of the propensity score has suggested that variables that are related to the outcome but not the exposure should be included in propensity score estimation because they improve the precision of the estimated TE without affecting bias,[71] as long as there is no unmeasured confounding and all confounders are controlled so that strong ignorability holds.[72-74] This same research showed that conversely, the inclusion of variables that are unrelated to outcome but related to exposure reduces the precision of the TE (without affecting bias). Our procedure will include variables related to the outcome but not the exposure (and our simulations do not include unmeasured confounding) and exclude variables unrelated to the outcome but related to the exposure, which may be a factor in this result. Future research should examine the case where not all confounders are measured.

Study 3 also showed that BART performed similarly to the synthetic RF estimator, with, in some cases, slightly lower bias (but higher RMSE). In moderate samples with a relatively moderate number of features, BART may be a reasonable alternative to RF. However, for very large problems (eg, those with extensive numbers of features, such as when genetic features are included), RF approaches will be much more scalable.

Finally, in study 4, we showed how the RF approach can be expanded to include survival outcomes and with multiple different treatments. The application for this examination used observational data. With a survival outcome, the estimate of ITEs is a function, and the ranking of treatments could thus change over time. Therefore, a particular treatment has a lower probability of early survival relative to a different treatment for an individual, but that same treatment may have higher rates of long-run survival; that is, the survival curves may cross. This issue will also arise when we consider longitudinal measures of outcome. There clearly may be treatment comparisons where outcomes are better immediately after a particular treatment; however, there is a quicker decay of the impact, so that in the longer run, the individual may be better off with the treatment that maintains its effectiveness longer. Our examinations have not addressed longitudinal outcomes (other than the survival context).

Study Limitations

Although our research has many strengths, there are some limitations. As noted previously, there are many machine learning approaches that could be used to estimate ITEs. Whereas we have expanded our investigation beyond what we proposed in our PCORI contract, there are still many approaches (eg, support vector machines and targeted learning) that we have not explored. Clearly, our research is limited by the focus on tree-based machine learning approaches. We feel this may be warranted due to their nonparametric basis, but other methods may have equal or better performance. Second, again as noted previously, we have not addressed ITE estimates when the outcome is longitudinally measured as in a trajectory of outcome after treatment. An important limitation is that we have not addressed individual uncertainty regarding the ITE estimate. In machine learning approaches where there is frequently an iterative procedure to select a model, this has been a rather difficult problem. We devoted considerable effort to using subsampling[61] of our entire PITE procedure, but we had little success. There have been very recent advances in this area on RFs. Wager and Athey[6] have shown that when trees are grown using subsampling without replacement, the results are asymptotically Gaussian, and this result can be used to form asymptotic confidence intervals. Their approach to estimating ITEs, however, estimates both treatments together and estimates treatment differences within the terminal node. This greatly simplifies the subsampling process.

Su et al[64] use the infinitesimal jackknife within their interaction tree approach. Loh et al[75] use a bootstrapping approach, whereas other researchers have used Bayesian approaches.[76,77] Finally, our assessment of confounding only explored the situation where all confounders are observed and included in the feature set. Exploration of situations where confounders are partially masked may be informative.

Future Research

On the methodological front, our study limitations could be used as a blueprint for future research. However, we feel that although there is more work to be done methodologically, these methods are at the point where increased emphasis should be on their application in specific fields. We are exploring ways to support greater application of these methods in the areas of cancer (Ishwaran and Lu) and in substance abuse treatment and behavioral treatments for HIV (Feaster).

CONCLUSIONS

RF is a flexible, nonparametric method that can be used to estimate ITEs, which could become an important component of treatment planning. It is one of several promising methods for estimating ITEs. When the relationship between an individual's characteristics and their outcomes under different treatments has simple, parametric forms and all the variables used to generate the data are included in the set of variables used to make the prediction, most of these methods, both RF based and many others, such as MI, can do a good job. However, when we created complex interaction terms (using transformations of variables that are then not included in the set of features used to predict outcomes), the MI method failed to make correct individual estimates.

We have extended the use of RF-based approaches to include observational confounded data. The RF approach works quite well at correcting confounding as long as all confounders are included in the set of characteristics used to predict the ITEs. We have also expanded these approaches to survival outcomes and multiple treatments. Our multiple-treatment approach specifically assesses whether the overlap in participant characteristics among the samples is sufficient to make pairwise comparisons.

REFERENCES

1. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med.* 2011;30(24):2867-2880.

2. Lamont A, Lyons MD, Jaki T, et al. Identification of predicted individual treatment effects in randomized clinical trials. *Stat Methods Med Res.* 2018;27(1):142-157.

3. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat.* 2011;20(1):217-240.

4. Tang F, Ishwaran H. Random forest missing data algorithms. *Stat Anal Data Min.* 2017;10(6):363-377.

5. Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat.* 2018;27(1):209-219.

6. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc.* 2018;113(523):1228-1242.

7. Kaplan SH, Billimek J, Sorkin DH, Ngo-Metzger Q, Greenfield S. Who can respond to treatment? Identifying patient characteristics related to heterogeneity of treatment effects. *Med Care.* 2010;48(6 ~~suppl~~Suppl):S9-S16.

8. Gibbons RD, Hur K, Brown CH, Davis JM, Mann JJ. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry.* 2012;69(6):572-579.

9. Materson BJ. Variability in response to antihypertensive drugs. *Am J Med.* 2007;120(4 ~~suppl~~Suppl 1):S10-S20.

10. Stroup TS. Heterogeneity of treatment effects in schizophrenia. *Am J Med.* 2007;120(4 ~~suppl~~Suppl 1):S26-S31.

11. Su X, Meneses K, McNees P, Johnson WO. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *J R Stat Soc Ser C Appl Stat.* 2011;60(3):457-474.

12. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)-uses of baseline data in clinical trials. *Lancet.* 2000;355(9209):1064-1069.

13. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med.* 2002;21(19):2917-2930.

14. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses: power and sample size for the interaction test. *J Clin Epidemiol.* 2004;57(3):229-236.

15. Cui L, James Hung H, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat.* 2002;12(3):347-358.

16. Lagakos SW. The challenge of subgroup analyses—reporting without distorting. *N Engl J Med.* 2006;354(16):1667-1669.

17. Peto R, Collins R, Gray R. Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol.* 1995;48(1):23-40.

18. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 2005;365(9454):176-186.

19. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med.* 2007;357(21):2189-2194.

20. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA.* 1991;266(1):93-98.

21. Alexander GC, Lambert BL. Is treatment heterogeneity an Achilles' heel for comparative effectiveness research? *Pharmacotherapy.* 2012;32(7):583-585.

22. PCORI. *The PCORI Methodology Report*. Posted November 2013. Accessed August 26, 2020. https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-report

23. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess (Rockv).* 2001;5(33):1-56.

24. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32.

25. Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics.* 2010;12(2):270-282.

26. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat.* 2013;7(1):443-470.

27. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med.* 2011;30(21):2601-2621.

28. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res.* 2009;10(Feb):141-158.

29. Doove LL, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classif.* 2014;8(4):403-425.

30. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat Med.* 2014;33(2):219-237.

31. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc.* 2005;100(469):322-331.

32. Metsch LR, Feaster DJ, Gooden L, et al. Effect of risk-reduction counseling with rapid HIV testing on risk of acquiring sexually transmitted infections: the AWARE randomized clinical trial. *JAMA.* 2013;310(16):1701-1710.

33. Jones RH, Velazquez EJ, Michler RE, et al. Coronary bypass surgery with or without surgical ventricular reconstruction. *N Engl J Med.* 2009;360(17):1705-1717.

34. Groenwold RH, Rovers MM, Lubsen J, van der Heijden GJ. Subgroup effects despite homogeneous heterogeneity test results. *BMC Med Res Methodol.* 2010;10(1):43. doi: 10.1186/1471-2288-10-43

35. Breiman L, Friedman JH, Olshen R, Stone C. *Classification and Regression Trees*. Chapman and Hall; 1984.

36. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat.* 1996;24(6):2350-2383.

37. Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol.* 2005;28(2):171-182.

38. Chen X, Liu C-T, Zhang M, Zhang H. A forest-based approach to identifying gene and gene–gene interactions. *Proc Natl Acad Sci U S A.* 2007;104(49):19199-19203.

39. Chen X, Wang L, Ishwaran H. An integrative pathway-based clinical–genomic model for cancer survival prediction. *Stat Probab Lett.* 2010;80(17):1313-1319.

40. Chen X-W, Liu M. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics.* 2005;21(24):4394-4400.

41. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics.* 2012;99(6):323-329.

42. Hsich E, Gorodeski E, Blackstone EH, Ishwaran H, Lauer M. Identifying important risk factors for survival in systolic heart failure patients using random survival forests. *Circ Cardiovasc Qual Outcomes.* 2011;4(1):39-45.

43. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC ~~Genetics~~Genet.* 2004;5(1):32. ~~doi:~~ doi:10.1186/1471-2156-5-32

44. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens.* 2005;26(1):217-222.

45. Rice TW, Rusch VW, Ishwaran H, Blackstone EH, Worldwide Esophageal Cancer Collaboration. Cancer of the esophagus and esophagogastric junction. *Cancer.* 2010;116(16):3763-3773.

46. Rizk NP, Ishwaran H, Rice TW, et al. Optimum lymphadenectomy for esophageal cancer. *Ann Surg.* 2010;251(1):46-50.

47. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci.* 2003;43(6):1947-1958.

48. Yoon DY, Smedira NG, Nowicki ER, et al. Decision support in surgical management of ischemic cardiomyopathy. *J Thorac Cardiovasc Surg*. 2010;139(2):283-293.

49. Wu B, Abbott T, Fishman D, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics.* 2003;19(13):1636-1643.

50. Ishwaran H, Malley JD. Synthetic learning machines. *BioData Min.* 2014;7(1):28. ~~doi:~~ doi:10.1186/s13040-0014-28-y

51. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41-55.

52. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med.* 2007;26(1):20-36.

53. Imai K, van Dyk DA. Causal inference with general treatment regimes. *J Am Stat Assoc.* 2004;99(467):854-866.

54. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010;29(3):337-346.

55. Ishwaran H. The effect of splitting on random forests. *Mach Learn.* 2015;99(1):75-118.

56. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Stat Comput.* 1999;9(2):123-143.

57. Dazard J-E, Rao JS. Local sparse bump hunting. *J Comput Graph Stat.* 2010;19(4):900-929.

58. Dazard J-E, Sunil Rao J, Markowitz S. Local sparse bump hunting reveals molecular heterogeneity of colon tumors. *Stat Med.* 2012;31(11-12):1203-1220.

59. Díaz-Pachón DA, Dazard J-E, Rao JS. Unsupervised bump hunting using principal components. In: Ahmed SE, ed. *Big and Complex Data Analysis: Methodologies and Applications.* Springer International Publishing; 2017:325-345.

60. Lu M, Blackstone E, Ishwaran H. Personalized treatment for ischemic cardiomyopathy: incorporating expert knowledge in random forest approaches using observational survival data with non-overlapping groups. Under review for publication as of August 26, 2020.

61. Politis DN, Romano JP, Wolf M. *Subsampling.* Springer Science & Business Media; 1999.

62. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer; 2009.

63. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1):Article25. https://doi.org/10.2202/1544-6115.1309

64. Su X, Peña AT, Liu L, Levine RA. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat Med.* 2018;37(17):2547-2560.

65. Xu Y, Yu M, Zhao YQ, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics.* 2015;71(3):645-653.

66. Deng A, Zhang P, Chen S, Kim DW, Lu J. Concise summarization of heterogeneous treatment effect using total variation regularized regression. Preprint. Posted October 13, 2016. arXiv 1610.03917. https://arxiv.org/abs/1610.03917

67. Zhao Q, Small DS, Ertefaie A. Selective inference for effect modification via the lasso. Preprint. Posted May 22, 2017. arXiv 1705.08020. https://arxiv.org/abs/1705.08020

68. Zhang Z, Seibold H, Vettore MV, Song W-J, François V. Subgroup identification in clinical trials: an overview of available methods and their implementations with R. *Ann Transl Med.* 2018;6(7):122. ~~doi:~~ doi:10.21037/atm.2018.03.07

69. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf.* 2012;21(~~suppl~~Suppl 2):138-147.

70. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *Am J Epidemiol.* 2011;174(5):613-620.

71. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163(12):1149-1156.

72. Wooldridge JM. Should instrumental variables be used as matching variables? *Res Econ.* 2016;70(2):232-237.

73. Bhattacharya J, Vogt WB. *Do Instrumental Variables Belong in Propensity Scores?* National Bureau of Economic Research Working Paper t0343. Published September 2007. Accessed June 26, 2020. https://prod.nber.org/papers/t0343

74. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol.* 2011;174(11):1213-1222.

75. Loh WY, Man M, Wang S. Subgroups from regression trees with adjustment for prognostic effects and postselection inference. *Stat Med.* 2019;38(4):545-557.

76. Alaa AM, van der Schaar M. Bayesian inference of individualized treatment effects using multi-task Ggaussian processes. Paper presented at: Advances in Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.

77. Hahn PR, Murray J, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Anal.* 2020; advance publication. https://projecteuclid.org/euclid.ba/1580461461

RELATED PUBLICATIONS

Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat.* 2018;27(1):209-219. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5920646/

Preprint version: https://arxiv.org/abs/1701.05306

Lu M, Blackstone E, Ishwaran H. Personalized treatment for ischemic cardiomyopathy: incorporating expert knowledge in random forest approaches using observational survival data with non-overlapping groups. 2017. Submitted to *Ann Appl Stat.*

ACKNOWLEDGMENTS

We would like to acknowledge our stakeholder advisory board: Allan Rodriguez, MD; Susanne Doblecki-Lewis, MD; David Forest, PhD; and Kira Villamizar, MPH.

Disclaimer:
The [views, statements, opinions] presented in this report are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee.

Acknowledgment:
Research reported in this report was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (#ME-1403-12907). Further information available at: https://www.pcori.org/research-results/2014/new-statistical-methods-assess-how-patients-different-traits-respond-same