BMC Medical Research Methodology

RESEARCH

Open Access

Assessing the properties of patient-specific treatment effect estimates from causal forest algorithms under essential heterogeneity

John M. Brooks[1,2*], Cole G. Chapman[3,4], Brian K. Chen[2,4], Sarah B. Floyd[4,5] and Neset Hikmet[4,6]

Abstract

Background Treatment variation from observational data has been used to estimate patient-specific treatment effects. *Causal Forest Algorithms (CFAs)* developed for this task have unknown properties when treatment effect heterogeneity from unmeasured patient factors influences treatment choice – *essential heterogeneity*.

Methods We simulated eleven populations with identical treatment effect distributions based on patient factors. The populations varied in the extent that treatment effect heterogeneity influenced treatment choice. We used the generalized random forest application (CFA-GRF) to estimate patient-specific treatment effects for each population. Average differences between true and estimated effects for patient subsets were evaluated.

Results CFA-GRF performed well across the population when treatment effect heterogeneity did not influence treatment choice. Under essential heterogeneity, however, CFA-GRF yielded treatment effect estimates that reflected true treatment effects only for treated patients and were on average greater than true treatment effects for untreated patients.

Conclusions Patient-specific estimates produced by CFAs are sensitive to *why* patients in real-world practice make different treatment choices. Researchers using CFAs should develop conceptual frameworks of treatment choice *prior to estimation* to guide estimate interpretation *ex post*.

Keywords Machine learning, Causal Forest Algorithm (CFA), Treatment effect estimation, Simulation modeling, Linear probability estimators

*Correspondence:
John M. Brooks
john-brooks@sc.edu
[1] Center for Effectiveness Research in Orthopaedics - Arnold School of Public Health Greenville, 915 Greene Street #302D, Columbia, SC 29208-0001, USA
[2] University of South Carolina Arnold School of Public Health, Health Services Policy & Management, Columbia, SC, USA
[3] Department of Pharmacy Practice and Science Iowa City, University of Iowa, Iowa, USA
[4] Center for Effectiveness Research in Orthopaedics, Greenville, SC, USA
[5] Clemson University College of Behavioral Social and Health Sciences, Public Health Sciences, Clemson, South Carolina, USA
[6] Department of Integrated Information Technology, Innovation Think Tank Lab @ USC, University of South Carolina College of Engineering and Computing, Columbia, SC, USA

Introduction

Developing *patient-specific treatment effect evidence* to guide individualized treatment decision-making is a cornerstone of patient-centered care [1–3]. The need for patient-specific evidence follows from the acknowledged breadth of outcome variation across patients receiving the same treatment. [4–10]. This phenomenon is known as *treatment effect heterogeneity* and is defined as "nonrandom variation in the direction of magnitude of a treatment effect" [11]. With their restrictive inclusion/exclusion criteria, randomized controlled trials cannot generate appropriate patient-specific evidence for many patients [4, 11–14]. As an alternative, observational data provide treatment variation within the context of real-world practice and a diversity of patients well beyond those evaluated in RCTs [2, 3, 12, 15, 16]. The traditional approach to estimate patient-specific treatment effects using observational data is to use parametric estimators and assign to each patient an estimated treatment effect from a "reference class" of patients [17–22]. Reference classes are defined a priori by the researcher based on combinations of measured patient factors that are conceptually associated with treatment effect heterogeneity [17–22]. The need to specify reference classes a priori has been described as "the central problem when using group evidence to forecast outcomes (or treatment effects) in individuals" [18]. Even with a small number of measured patient factors, a patient could be placed in many reference classes, leaving it unclear which class is best aligned to the patient [10, 17, 18].

Causal forest algorithms (CFAs) have been proposed to estimate patient-specific treatment effects in a manner that essentially assigns patients to reference classes *ex post* using information from the data, thereby eliminating the need to assign patients to reference classes a priori [23–33]. Simulation modeling has shown that CFAs can accurately estimate patient-specific treatment effects in scenarios in which treatment effect heterogeneity does not influence treatment choice [24, 26–29, 34–37]. However, in many real-world scenarios it is conceivable that unmeasured patient factors associated with treatment effectiveness influence treatment choice. This is called *essential heterogeneity* or sorting on the gain in the econometrics literature [38–51]. The properties of parametric treatment effect estimators under essential heterogeneity are well known [38–51]. However, the impact of essential heterogeneity on patient-specific treatment effect estimates using CFAs has not been evaluated. In this paper, we contrast the properties of patient-specific treatment effect estimates using the causal forest algorithm within the generalized random forests application (CFA-GRF) across simulation scenarios that vary in the extent that unmeasured patient factors associated with treatment effectiveness influence treatment choice.

Methodological background

Assigning patients into appropriate reference classes using observational data either a priori with parametric estimators or *ex post* through a CFA does not ensure that the resulting treatment effect estimates are appropriate for each patient. The conventional criticism of using observational data to estimate treatment effects is the risk of omitted variable bias in which unmeasured factors with direct effects on study outcomes are distributed differently between treated and untreated patients [52]. However, even if patients were assigned to appropriate reference classes and omitted variable bias risk is mitigated through study design, a single treatment effect estimate for a reference class may not be appropriate for each patient within a class. The econometric literature has shown that parametric estimators yield average treatment effect estimates for patient subsets based on treatment choice [38–67]. Under the assumption of no omitted variable bias, regression-based estimators yield unbiased estimates of the average treatment effect for the subset patients who chose treatment or the *average treatment effect on the treated* (ATT) [43, 48–50, 54, 57, 60, 68, 69]. Consequently, if treatment choice in an empirical setting was influenced by unmeasured patient factors related to treatment effectiveness – *essential heterogeneity* – the parametric estimate of ATT for a reference class will overstate the true treatment effects for the untreated patients in the class [39, 49, 50, 70]. Researchers using parametric estimators have learned not to generalize a single parametric treatment effect estimate to all patients in a population [38, 43, 47–51, 53, 55, 56, 58, 59, 61, 67, 70, 71].

In contrast, the properties of estimated patient-specific treatment effects from CFAs under essential heterogeneity have not been explored. Simulation research has demonstrated that CFAs accurately yield patient-specific treatment effects under the broad condition of *ignorability* [24, 26–29, 34–36]. Ignorability assumes that omitted variable bias does not exist within an empirical setting. However, ignorability also assumes that essential heterogeneity does not exist. These dual assumptions can be described using potential outcome notation. Define Y_{1i} and Y_{0i} as the potential outcomes for patient "i" when treated and untreated, respectively, and $(Y_{1i} - Y_{0i})$ is the true potential treatment effect for patient "i". Define T_i as the observed treatment choice for patient "i" and X_i as the set of measured patient factors available to the researcher. Ignorability is broadly defined as $(Y_{1i}, Y_{0i}) \perp T_i \mid X_i$ or conditional on X_i, treatment choice is independent of *both* potential patient outcomes [72]. As such, ignorability implies the following two distinct assumptions.

$$(Y_{0i}) \perp T_i | X_i \tag{1.1}$$

Assumption (I.1) says that, within a reference class of patients based on X_i, treatment choice is unrelated to *untreated potential outcomes* across patients. Or stated differently, treatment choice is unrelated to unmeasured patient factors associated with Y_{0i}. Assuming (I.1) eliminates the risk of omitted variable bias in an observational study [52].

Even if assumption (I.1) is true though, treatment effects may remain heterogeneous within a reference class defined by X_i. With respect to this heterogeneity, ignorability further assumes:

$$(Y_{1i} - Y_{01}) \perp T_i | X_i \tag{1.2}$$

Assumption (I.2) says that, within a reference class of patients defined by X_i, treatment choice within the class is *not* influenced by unmeasured patient factors associated with treatment effectiveness or there is no *essential heterogeneity* [38, 39, 45]. If ignorability holds within a reference class defined by X_i, only the treatment variation that stems from patient factors *unrelated to treatment effectiveness* will be used to estimate treatment effects within the class. Consequently, CFA simulation results which assume ignorability provide no guidance on the properties of patient-specific treatment effect estimates in real-world scenarios in which essential heterogeneity is thought to exist *a priori*. For example, the effectiveness of surgery for patients with shoulder fractures is thought to vary with fracture complexity and patient resiliency, which in turn influence surgery choice [73–77], but fracture complexity and patient resiliency are not measurable in large observational databases such as Medicare claims data [73–77]. A study using a causal forest algorithm to estimate patient-specific surgery effects using Medicare claims data theorized a priori that the resulting estimates should be interpreted in terms of essential heterogeneity, but evidence was not available to guide these interpretations [78]. In addition, understanding influence of essential heterogeneity on CFA estimates is especially relevant to researchers proposing to use CFAs in *effectiveness-implementation hybrid study designs* in which the *promotion* of a treatment is randomized to satisfy assumption (I.1) but decision makers still have the discretion to choose among available treatments based on individual patient factors [79–95].

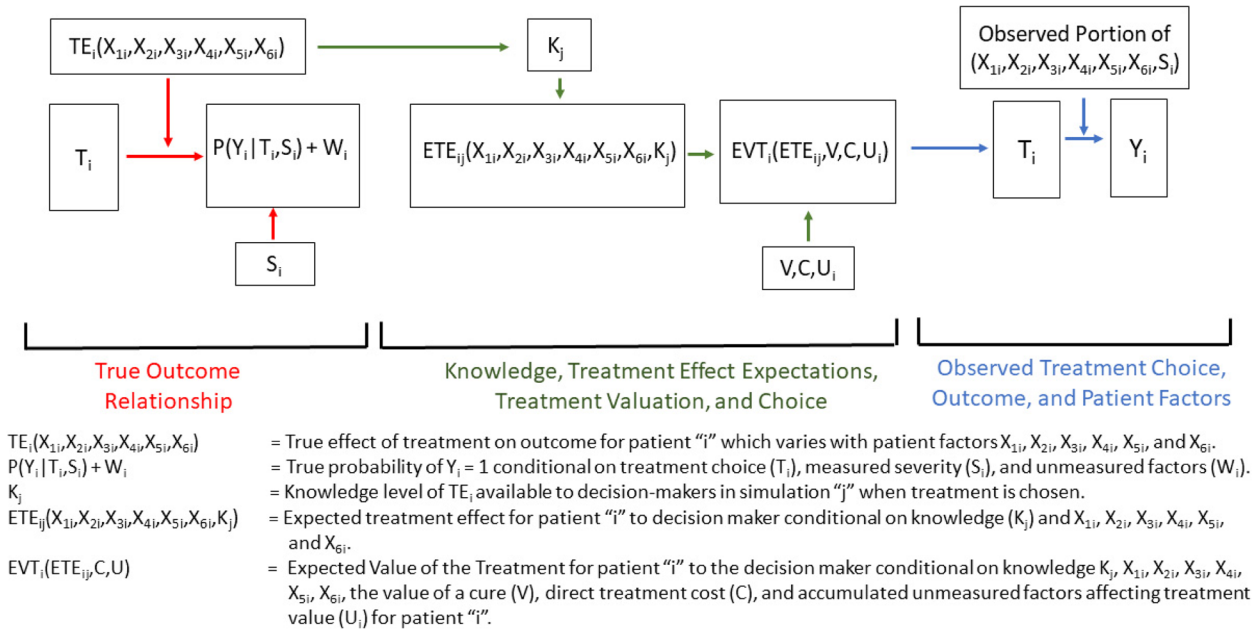
To provide this guidance, this study modified a treatment choice-based simulation method used in previous research to assess the impact of essential heterogeneity on patient-specific treatment effect estimates from a CFA estimator [43, 48, 53]. Eleven patient populations were simulated with the same distribution of true treatment effects drawn from identical distributions of simulated patient factors. All eleven simulations were specified to satisfy assumption (I.1). The simulations varied by plausible differences in the extent to which knowledge of true patient-specific treatment effects influenced treatment choice. We used the causal forest algorithm within the generalized random forests application (CFA-GRF) [24–26, 96, 97] to estimate patient-specific treatment effects for each simulated population. CFA-GRF has been singled out as the most appropriate CFA for estimating patient-specific treatment effects [98]. To tease out the influence of essential heterogeneity, we applied CFA-GRF to each simulated population under conditions of (1) *fully observed heterogeneity* in which all patient factors associated with treatment effect heterogeneity are observed by the researcher and (2) *partially observed heterogeneity* in which only a subset of the patient factors associated with treatment effect heterogeneity are observed by the researcher. Patient-specific treatment effect estimates from CFA-GRF were used to calculate the average absolute and average percentage differences between true and estimated effects for each simulated population and for treatment choice-based population subsets.

Methods

Simulation model

Our simulation model follows the general framework in the essential heterogeneity literature [39, 43, 45, 48, 53, 99]. Figure 1 contains a directed acyclic graph (DAG) illustrating the conceptual framework of treatment effect heterogeneity, treatment choice, and outcome within our simulations. Figure 1 was adapted from standard DAG approaches to reflect patient factors affecting treatment effectiveness and the treatment effect knowledge of the decision maker [100, 101]. Outcome (Y_i) equals 1 if patient "i" is cured of the medical condition, and 0 if not cured. $P(Y_i|T_i,S_i)$ is the probability of cure for patient "i" conditional on treatment choice (T_i) and patient severity (S_i). Patient cure probability also varies with accumulated other factors (W_i). Treatment (T_i) equals 1 if the patient receives treatment and 0 otherwise, which we designate as *watchful waiting*. In all simulations, the true absolute treatment effect for each patient "i" (TE_i) on Y_i relative to watchful waiting varies with six factors X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i} based on the following equation:

$$TE_i\left(X_{1i,}X_{2i}X_{3i}X_{4i}X_{5i}X_{6i}\right) = \beta_1{*}X_{1i} + \beta_2{*}X_{2i} + \beta_3{*}X_{3i} + \beta_4{*}X_{4i} + \beta_5{*}X_{5i} + \beta_6{*}X_{6i} \tag{1}$$



$TE_i(X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i})$ = True effect of treatment on outcome for patient "i" which varies with patient factors X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i}.
$P(Y_i|T_i,S_i)+W_i$ = True probability of $Y_i = 1$ conditional on treatment choice (T_i), measured severity (S_i), and unmeasured factors (W_i).
K_j = Knowledge level of TE_i available to decision-makers in simulation "j" when treatment is chosen.
$ETE_{ij}(X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i},K_j)$ = Expected treatment effect for patient "i" to decision maker conditional on knowledge (K_j) and X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i}.
$EVT_i(ETE_{ij},C,U)$ = Expected Value of the Treatment for patient "i" to the decision maker conditional on knowledge K_j, X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, the value of a cure (V), direct treatment cost (C), and accumulated unmeasured factors affecting treatment value (U_i) for patient "i".

Fig. 1 Directed Acyclic Graph (DAG) Describing the Conceptual Framework for the Simulation Model in which Patient Factors Affecting Treatment Effectiveness Affect Treatment Choice through Decision Maker Knowledge

X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i} are binary variables distributed Bernoulli for each patient with a probability of 0.5. Each β_x equals the absolute change in treatment effect if a patient has condition "X" ($\beta_1 = 0.024$, $\beta_2 = 0.048$, $\beta_3 = 0.071$, $\beta_4 = 0.095$, $\beta_5 = 0.119$, $\beta_6 = 0.143$). With these parameter values, simulated patients have true treatment effects ranging from 0 to 0.5 with an average true treatment effect of 0.25 for each simulated population. For example, if the simulated patient factors for patient "i" ($X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i}$) were (1,0,1,0,1,0), then patient "i's" true TE_i was.214 = (0.024 + 0 + 0.071 + 0 + 0.095 + 0). Figure 2 illustrates the identical distribution of

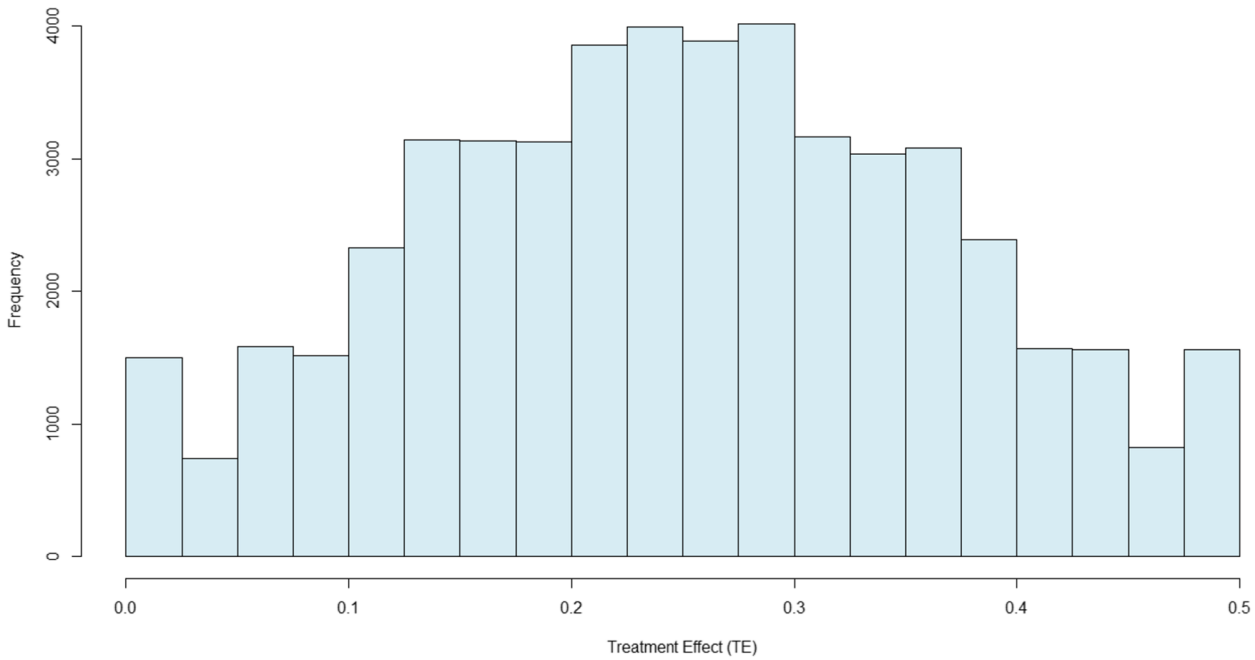


Fig. 2 Distribution of True Absolute Treatment Effects (TEi) Used in All Eleven Simulated Populations

simulated treatment effects across all eleven simulations in this study.

The true cure probability relationship for each simulated patient "i" signified by the red arrows in Fig. 1 is as follows:

$$\text{Probability of } Y_i = P(Y_i|T_i, S_i) + W_i = (\alpha_O + \alpha_S \cdot S_i + TE_i(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}) \cdot T_i) + W_i \quad (2)$$

α_0 equals the untreated patient cure probability at the mean severity level and was set to 0.1 in all simulations. Patient severity (S_i) was specified as a uniformly distributed random variable from -0.5 to 0.5. α_S equals the change in untreated patient cure probability for differences in severity level and was set to -0.1 in all simulations. As a result, in each simulated population, watchful waiting patients ($T_i = 0$) had a cure probability ranging from 0.05 to 0.15. Treated patients ($T_i = 1$) had a cure probability ranging from 0.05 to 0.65. All other unmeasured patient factors impacting the probability of a cure are found in (W_i).

The green arrows in Fig. 1 describe the treatment choice process that varied across the eleven simulations. In each simulation, it is assumed that the treatment decision-maker observes X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i} and forms an expected treatment effect for patient "i". The simulations differ by the *knowledge* available to decision makers of the relationship between the six patient factors and treatment effectiveness, as represented by the expected treatment effect function for simulation "j":

$$ETE_{ij}(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}K_j) = K_j * (TE_i(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}) - .25) + .25. \quad (3)$$

$K_j \in (0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1)$ is the proportion of patient-specific TE_i knowledge used by decision makers in simulation "j" that is distinct from the average population treatment effect. Decision makers are more aware of each patient's true treatment effect relative to the average population treatment effect as K_j increases from 0 to 1 across simulations. For example, in the simulation in which $K_j = 0$, decision makers only have knowledge of the average treatment effect across the population (0.25) when making treatment decisions for each patient. Alternatively, when $K_j = 1$, decision makers have exact knowledge of the treatment effect for patient "i" from observed X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i}. $ETE_{ij}(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, K_j)$ is used to calculate the expected value of treatment for patient "i" based on the following:

$$EVT_i(ETE_{ij}, V, C, U_i) = V \cdot ETE_{ij}(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, K_j) - C + U_i \quad (4)$$

$EVT_i(ETE_{ij},V,C,U_i)$ sums the expected benefits and detriments (e.g., costs) of treatment relative to watchful waiting for patient "i" that is conditional on knowledge K_i, X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i}, direct treatment cost C, cure value V, and U_i other accumulated factors affecting treatment value, which are independent of treatment effectiveness for patient "i". $ETE_{ij}(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, K_j)$ equals the decision maker's expected change in cure probability from treatment. To focus this study on the impact of essential heterogeneity across simulations, all patients were assigned a cure value V of \$800 and a treatment cost C of \$200. These values were chosen because they yield simulated population treatment percentages of approximately 50%. V designations of \$500 and \$1100 were also tried, which yielded different population treatment percentages but did not influence the interpretation of our results relative to the essential heterogeneity. U_i is the source of treatment valuation that varies across patients, is unrelated to treatment effectiveness and is unmeasured by the researcher. U_i values were assigned to patients from a normal distribution with a mean of zero and a common variance σ_U^2 across simulations. Furthermore, in all simulations, U_i was specified independently of W_i so that the differences in unmeasured factors influencing treatment choice had no relationship with the unmeasured factors directly effecting cure so that ignorability assumption (I.1) was satisfied.

In all simulations, decision makers chose treatment for patient "i" if EVT_i was positive and watchful waiting if EVT_i was negative. In the simulation in which the knowledge of patient-specific treatment effect heterogeneity is zero ($K_j = 0$), only variation in U_i leads to different treatment choices across simulated patients. As K_j increases across simulations, a larger proportion of the variation in treatment choice variation is attributable to treatment effectiveness or *sorting on the gain*. Once a treatment was chosen for each patient, cure (Y_i) was simulated using a Bernoulli function of $P(Y_i|T,S_i)$ for patient "i", given T_i and S_i. Table 1 summarizes the model parameters and values used in the simulations.

Table 1 Summary of simulation model parameters

Parameter	Description	Value and Distribution
β_1	Absolute increase in treatment effect on cure when $X_1 = 1$	.024
β_2	Absolute increase in treatment effect on cure when $X_2 = 1$	.048
β_3	Absolute increase in treatment effect on cure when $X_3 = 1$	.071
β_4	Absolute increase in treatment effect on cure when $X_4 = 1$	.095
β_5	Absolute increase in treatment effect on cure when $X_5 = 1$	.119
β_6	Absolute increase in treatment effect on cure when $X_6 = 1$	.143
TE_i	True treatment effect on outcome for patient "i" as a function of X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i}	Ranges from 0 to .5. Distribution in Fig. 2
S_i	Patient "i" severity level directly effecting cure but have no effect on treatment effectiveness and are unrelated to treatment choice	Distributed Uniform(-.5,.5)
α_0	Untreated patient cure probability at mean severity level	.1
α_S	Change in untreated patient cure probability given S_i	-.1
V	The value patients gain when cured	$800
C	The cost of treatment	$200
K_j	The proportion of knowledge of treatment effectiveness that is patient-specific in simulation "j"	$\begin{pmatrix} 0, .1, .2, .3, .4, .5, .6, \\ .7, .8, .9, 1 \end{pmatrix}$
U_i	Accumulated unmeasured factors for patient "i" which affect treatment valuation	N(0,25)
ETE_i	Expected treatment effect for patient "i" given knowledge within simulation "j"	$K_j * (TE_i — .25) + .25$
EVT_i	Expected value of treatment for patient "i" given ETE_i	$V \bullet ETE_i + C + U_i$
T_i	1 if patient is EVE_i is greater than 1, 0 otherwise	
$P(Y_i\|T_i,S_i)$	Probability patient "i" is cured given TE_i, T_i, and S_i	$.1 + TE_i \bullet T_i + (-.1) \bullet S_i$
Y_i	1 if patient is cured, 0 otherwise	Bernoulli function of $P(Y_i\|T_i,S_i)$
W_i	Unmeasured patient factors causing variation in Y_i given T_i and S_i	

To support large sample properties, we generated 50,000 patients in each simulation. The blue arrows in Fig. 1 describe the variables observed by the researcher after each simulation. By varying the knowledge of TE_i across simulations with K_j and the patient factors observed by the researcher, we can tease out the impacts of essential heterogeneity on patient-specific treatment effect estimates. In each scenario, researchers observe T_i, Y_i, S_i. We designate "fully observed heterogeneity" as the empirical condition in which researchers observe all six patient factors X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, and X_{6i}. We designate "partially observed heterogeneity" as the empirical condition in which researchers observe only X_{1i}, X_{2i}, X_{3i}, and X_{4i}. Under fully observed heterogeneity, treatment effects are homogeneous within each reference class spanned by combinations of the complete set of patient factors. When $K_j = 0$, decision-makers are not knowledgeable of the sources of treatment effect heterogeneity, and treatment choice varies only with U_i. Under fully observed heterogeneity with $K_j > 0$, decision-makers are at least partly knowledgeable of the sources of treatment effect heterogeneity, with the effect of this knowledge on treatment choice increasing with K_j. Under partially observed heterogeneity, treatment effects are heterogeneous within the reference classes defined by the observed set of patient factors. Partially observed heterogeneity with $K_j = 0$ has been dubbed *nonessential heterogeneity* in the econometric literature [38, 39]. Under nonessential heterogeneity, treatment choice is not influenced by the unmeasured patient factors affecting treatment effectiveness within a reference class. Scenarios with partially observed heterogeneity and $K_j > 0$ represent *essential heterogeneity*. In these scenarios, treatment effects are heterogeneous within each reference class, with the influence of treatment effect heterogeneity on treatment choice increasing with K_j across simulations.

Estimation methods

Simulated population summaries

Treatment effect estimation using observational data requires what is called a common area of support or overlap between treated and untreated patients or that patients with the same measured patient factors must be observed to make different treatment choices [102, 103]. It has been shown that including patients in study populations with insufficient overlap can lead to biased treatment effect estimates [104, 105]. The treatment choice-based simulations used here naturally reduce overlap the more that treatment choice is influenced by patient factors affecting treatment effectiveness. To monitor this influence across simulations, we used the

SAS PROC LOGISTIC procedure to estimate the treatment propensity score for each patient in each simulated population under both "fully observed heterogeneity" and "partially observed heterogeneity". Each simulated patient was then designated into either the "overlapped" subset with a propensity score between 0.05 and 0.95 or into the nonoverlapped subset with propensity scores either less than 0.05 or greater than 0.95 [104, 105]. We then estimated the percentage of patients in each simulated population who were treated, untreated, overlapped and treated, overlapped and untreated, nonoverlapped and treated, and nonoverlapped and untreated and then calculated the true average TE_i in each subset.

Next, for each simulated population, we estimated a linear probability model (LPM) of treatment choice T_i on true TE_i using the SAS PROC REG procedure with the SCORR1 option. This procedure provides the percentage of treatment choice variation within the simulated population that is attributable to variation in the true treatment effect to serve as a measure of the influence of the true treatment effect on treatment choice. Last, we estimated the effect of T_i and S_i on Y_i using a LPM in each simulated population. The parametric treatment effect literature states that the LPM estimator of the parameter on T_i will yield a consistent estimate of the average absolute treatment effect on the treated in each simulated population [43, 48–50, 54, 57, 60, 68, 69].

Casual forest algorithm

We then applied the CFA-GRF [24–26, 96, 97] using the "grf" package in R [106] to estimate treatment effects for each patient in each simulated population. CFA-GRF evolved from standard classification and regression tree (CART) and random forest ensemble methods [24–26, 96, 97]. CART procedures iteratively partition "nodes" of observations within a population into subnodes or "branches" based on measured factors in a manner that maximizes the differences in an outcome across possible branches [97]. A tree is formed by viewing all of the subsequent branches of the study population. The final subnode or leaf on the end of a branch can be thought of as an algorithm-generated *ex post* reference class for observations with factors matching the leaf. The random forest approach is an ensemble method that generates a "forest" of CART trees through resampling from the study population [96]. The estimated outcome for a single observation is the average outcome across the leaves in the trees in the forest containing that observation. CFA-GRF extends the random forest approach to the goal of estimating the causal effect of a predictor of interest (e.g., a treatment) on an outcome. CFA-GRF partitions observations based on measured factors in a manner that maximizes the expected differences *in the estimated treatment effect* on an outcome [24–26]. For each simulated population, CFA-GRF was run using 4000 trees, minimum leaf sizes of 50 and the "honest" approach suggested by the algorithm creators, in which trees were estimated using a randomly selected 25% of the simulated population [26]. We ran CFA-GRF specifying X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}, and S_i in the "fully observed heterogeneity" specification and X_{1i}, X_{2i}, X_{3i}, X_{4i}, and S_i in the "partially observed heterogeneity" specification. As a result, each patient in each simulated population had two treatment effect estimates. We assessed the properties of these estimates by evaluating their ability to identify average treatment effect parameters for each simulated population and treatment choice-based subsets of the population. We calculated the average absolute and percentage difference between the true treatment effect for each simulated patient (TE_i) and estimated treatment effects for the full population and subsets of population based on treatment choice and propensity score "overlap" status.

Results

Summary information across simulated populations

Table 2 summarizes each simulated population. Column A in Table 2 shows the proportion of treatment effect expectations (ETE_i) shaped by the true effect for each patient (TE_i) in each simulation – K_j from Eq. (3). Column B shows the percentage of treatment choice variation in each simulation explained by TE_i. Columns C and D show the percentage of simulated patients who *overlapped* or had propensity scores greater than 0.05 and less than 0.95 in the fully observed heterogeneity and partially observed heterogeneity scenarios, respectively. Columns E through J show the true average TE_i for subsets of treated, untreated, overlapped and treated, overlapped and untreated, nonoverlapped and treated, and nonoverlapped and untreated patients, respectively. These columns also show in parentheses the percentage of patients within each subset.

Patient-specific treatment effects (TE_i) do not influence treatment choice in simulation 1, and as a result, the average true TE_i is close to the true population average treatment effect of 0.25 for both treated and untreated patients. Moving from simulations 2 through 11, though, the knowledge of TE_i increases in decision making, and TE_i explains a larger portion of the variation in treatment choice (column B). Under fully observed heterogeneity, all patients are fully overlapped in simulations 1 through 6. The percentage of overlapping patients falls from 97.0% to 68.8% in simulations 7 through 11. Under the partially observed heterogeneity, all patients overlapped across all simulations. Columns E and F show how the greater influence of TE_i on treatment choice leads to sorting on the gain. The average TE_i for the treated patients in Column

Table 2 Summary information for simulated populations

Simulation	A	B	C	D	E	F	G	H	I	J	K
	Proportion of true (TE_i) influencing expect treatment effect (ETE_i) – (K_i)[a]	% of Treatment Choice Variation Explained by (TE_i)[b]	Fully Observed Heterogeneity: % of Patients Overlapped[c]	Partially Observed Heterogeneity: % of Patients Overlapped[d]	Average True Absolute Treatment Effect (TE_i) Within Subset (Percentage of Patients)						Parametric Linear Probability Model Estimate (ATT)
					Full Population		Overlapped with Fully Observed Heterogeneity		Non-Overlapped with Fully Observed Heterogeneity		
					Treated	Untreated	Treated	Untreated	Treated	Untreated	
1	0	.0006	100	100	.250 (49.8)	.251 (50.2)	.250 (49.8)	.251 (50.2)			.249
2	.10	.18	100	100	.256 (49.8)	.246 (50.2)	.256 (49.8)	.246 (50.2)			.255
3	.20	1.4	100	100	.264 (50.0)	.237 (50.0)	.264 (50.0)	.237 (50.0)			.267
4	.30	5.3	100	100	.277 (50.1)	.224 (49.9)	.277 (50.1)	.224 (49.9)			.276
5	.40	11.9	100	100	.290 (50.1)	.212 (49.9)	.290 (50.1)	.212 (49.9)			.292
6	.50	20.1	100	100	.301 (50.2)	.200 (49.8)	.301 (50.2)	.200 (49.8)			.300
7	.60	27.8	97.0	100	.310 (50.2)	.191 (49.8)	.304 (48.7)	.196 (48.3)	.500 (1.5)	.001 (1.5)	.307
8	.70	34.5	90.7	100	.317 (50.3)	.184 (49.7)	.300 (45.5)	.200 (45.2)	.475 (4.8)	.025 (4.5)	.316
9	.80	39.8	84.5	100	.322 (50.2)	.179 (49.8)	.297 (42.3)	.203 (42.1)	.457 (7.9)	.044 (7.6)	.321
10	.90	44.3	78.3	100	.326 (50.2)	.175 (49.8)	.293 (39.2)	.207 (39.1)	.442 (11.0)	.059 (10.7)	.321
11	1.00	48.0	68.8	100	.329 (50.3)	.172 (49.7)	.285 (34.4)	.214 (34.4)	.423 (15.8)	.077 (15.3)	.329

[a] The proportion of patient-specific TE_i knowledge used by decision makers in simulation "j" in developing the expected treatment effect for patient "i" that is distinct from the population average treatment effect based on the equation $ETE_i = K_j * (TE_i(X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i})—.25) + .25$. The population average treatment effect is .25 in all simulations

[b] The percentage of treatment choice variation explained by TE_i using a linear probability model of treatment choice T_i on true TE_i using SAS PROC REG procedure with the SCORR1 option

[c] Percentage of patients in sample with treatment propensity score greater than .05 and less than .95 when all six patient factors are fully specified in the propensity score equation

[d] Percentage of patients in sample with treatment propensity score greater than .05 and less than .95 when only X_{1i}, X_{2i}, X_{3i}, X_{4i} factors are specified in the propensity score equation

E increased from 0.250 to 0.329 as K increased from 0 to 1, while the average TE_i for the untreated patients in Column F fell from 0.251 to 0.172 across this range. Columns G through J stratify treated and untreated patients by overlap status under fully observed heterogeneity. The average TE_i of nonoverlapped treated patients (column I) is greater than that of overlapped treated patients (column G). Likewise, the average TE_i of nonoverlapping untreated patients (column J) is less than that of overlapping untreated patients (column H). Column K of Table 2 shows the estimated treatment effect for the full population in each simulation using a linear probability model (LPM). A comparison of these estimates with column E confirms that LPM yields estimates of the average treatment effect on the treated (ATT) [57]. When treatment effects are heterogeneous, LPM estimates appropriately generalize to untreated patients only when TE_i does not influence treatment choice, as in simulation 1 [57].

CFA-GRF results under fully observed heterogeneity

Table 3 contains the average percentage differences between the true treatment effects and individual treatment effect estimates from CFA-GRF for each of the eleven simulated populations under fully observed heterogeneity. Estimates are reported for the full population in each simulation and treatment-choice-based subsets. Table A.1 in the Additional file 1 shows these results in terms of average absolute differences between the true treatment effect values and estimated treatment effects. The percentage differences in Table 3 were calculated using the average true treatment effect for each population subset found in Table 2 and the average absolute differences for each subset in Table A.1. For example, the average percentage difference between the estimated and true treatment effect values for the full population in simulation 1 under fully observed heterogeneity is 100*(-0.0014)/0.25 = -0.56%. Column E of Table 3 shows that under fully observed heterogeneity on average, CFA-GRF produces treatment effect estimates that reflect each population across simulations. However, as treatment choice becomes more responsive to TE_i, CFA-GRF estimates increasingly understate the true treatment effect for treated patients and overstate the true treatment effect for untreated patients. Simulation 1 under fully observed heterogeneity fully satisfies ignorability, and CFA-GRF produces patient-specific treatment effect estimates that on average reflect the true patient treatment effects for the entire population and for both treated and untreated patient subsets. In contrast, in simulation 11, in which decision-makers have full knowledge of TE_i in treatment choice, the treatment effect estimates for treated patients are on average 14.74% lower than the truth, and the estimated treatment effects for untreated patients are on average 30.99% higher than the truth. These percentage differences are not symmetric because untreated patients have a lower average true treatment effect. Columns G to J in simulations 6 through 11 demonstrate that these differences exist for both overlapping and nonoverlapping patients but are more pronounced for nonoverlapping patients.

CFA-GRF results under partially observed heterogeneity

Table 4 contains the average percentage differences between the true treatment effect values and CFA-GRF treatment effect estimates for each simulated population under partially observed heterogeneity. Under partially observed heterogeneity all patients are overlapped so that the columns G through J found in Table 3 are unnecessary. Under ignorability in simulation 1, CFA-GRF again produces estimates that on average are close to true patient treatment effects for the entire population and for the treated and untreated patient subsets. In simulation 1, CFA-GRF estimates under partially observed heterogeneity had larger standard errors than those under fully observed heterogeneity (see Table A.2). Treatment effects estimated from CFA-GRF for treated patients closely reflect their true values across all eleven simulations. In contrast, CFA-GRF estimates for untreated patients are higher than their true values across simulations 2 through 11, with the differences increasing with the level of TE_i influence on treatment choice. For example, based on the true average treatment effect for untreated patients from Table 2 and the average absolute differences for each population in Table A.1, on average, CFA-GRF estimates for untreated patients are 2.4% greater than their true values in simulation 2 – 100*(0.006)/(0.246)) and 76.3% greater than their true values in simulation 11 – 100*(0.1312)/(0.172). As a result, when TE_i influences treatment choice under partially observed heterogeneity, CFA-GRF estimated treatment effects across the whole population are on average greater than their true values.

Discussion

Causal forest algorithms (CFAs) have been proposed to estimate patient-specific treatment effect evidence using observational data [23–33, 107]. To apply CFAs, observational databases must contain patients with similar combinations of measured factors who were observed to make different treatment choices. The positive properties of CFAs for estimating patient-specific treatment effects have been established using simulation models under the assumption of ignorability [26–29, 34–36]. Under ignorability, only the treatment variation from *unobserved patient factors not associated with treatment effect heterogeneity* is available to estimate patient-specific treatment effects. Therefore, it is unknown whether the

Table 3 Average Percentage Differences Between the Estimated Treatment Effects and True Treatment Effects from the Causal Forest Algorithm within the Generalized Random Forests Application (CFA-GRF) Under *Fully Observed Heterogeneity* Across Simulated Populations Which Differ by the Extent That Treatment Effect Influences Treatment Choice

Simulation	A	B	C	D	E	F	G	H	I	J
	Proportion of true (TE_i) influencing (ETE_i) at Treatment Choice – (K_j)[a]	% of Treatment Choice Variation Explained by (TE_i)[b]	% of Patients Overlapped[c]	Average Percentage Difference Between True and Estimated Treatment Effects						
				Full Population	Treated	Untreated	Overlapped with Fully Observed Heterogeneity		Non-Overlapped with Fully Observed Heterogeneity	
							Treated	Untreated	Treated	Untreated
1	0	.0006	100	-0.56%	-0.64%	-0.44%				
2	.10	.18	100	-0.36%	-0.39%	-0.28%				
3	.20	1.4	100	0.56%	0.38%	0.76%				
4	.30	5.3	100	-0.84%	-1.12%	-0.54%				
5	.40	11.9	100	0.48%	-1.21%	2.83%				
6	.50	20.1	100	-0.24%	-3.55%	4.75%				
7	.60	27.8	97.0	2.76%	-1.84%	10.31%	-1.15%	8.62%	-14.98%	11110.00%
8	.70	34.5	90.7	-0.96%	-9.78%	14.51%	-7.07%	8.35%	-26.27%	508.00%
9	.80	39.8	84.5	2.84%	-7.74%	22.07%	-3.13%	11.48%	-23.85%	293.41%
10	.90	44.3	78.3	0.08%	-11.66%	22.11%	-5.53%	8.79%	-26.11%	192.88%
11	1.00	48.0	68.8	0.84%	-14.74%	30.99%	-5.33%	10.47%	-28.58%	159.22%

[a] The proportion of patient-specific TE_i knowledge used by decision makers in simulation "j" in developing the expected treatment effect for patient "i" that is distinct from the population average treatment effect based on the equation $ETE_i = K_j * (TE_i(X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i}) — .25) + .25$

[b] The percentage of treatment choice variation explained by TE_i using a linear probability model of treatment choice T_i on true TE_i using SAS PROC REG procedure with the SCORR1 option

[c] Percentage of patients in sample with treatment propensity score greater than .05 and less than .95 when all six patient factors are fully specified in the propensity score equation

Table 4 Average Percentage Differences Between the Estimated Treatment Effects and True Treatment Effects from the Causal Forest Algorithm within the Generalized Random Forests Application (CFA-GRF) Under *Partially Observed Heterogeneity* Across Simulated Populations Which Differ by the Extent That Treatment Effect Influences Treatment Choice

Simulation	A	B	C	D	E	F
	Proportion of true (TE_i) influencing (ETE_i) at Treatment Choice – (K_j)[a]	% of Treatment Choice Variation Explained by (TE_i)[b]	% of Patients Overlapped[c]	Average Percentage Difference Between True and Estimated Treatment Effects		
				Full Population	Treated	Untreated
1	0	.0006	100	-1.16%	-1.08%	-1.27%
2	.10	.18	100	1.12%	-0.20%	2.44%
3	.20	1.4	100	4.12%	0.57%	8.02%
4	.30	5.3	100	6.44%	-0.36%	14.87%
5	.40	11.9	100	12.12%	0.72%	27.69%
6	.50	20.1	100	15.08%	0.27%	37.40%
7	.60	27.8	100	18.68%	0.58%	48.22%
8	.70	34.5	100	18.60%	-1.51%	53.42%
9	.80	39.8	100	24.36%	1.27%	66.03%
10	.90	44.3	100	22.00%	-1.75%	66.51%
11	1.00	48.0	100	25.44%	-1.03%	76.28%

[a] The proportion of patient-specific TE_i knowledge used by decision makers in simulation "j" in developing the expected treatment effect for patient "i" that is distinct from the population average treatment effect based on the equation $ETE_i = K_j * (TE_i(X_{1i},X_{2i},X_{3i},X_{4i},X_{5i},X_{6i})—.25) + .25$

[b] The percentage of treatment choice variation explained by TE_i using a linear probability model of treatment choice T_i on true TE_i using SAS PROC REG procedure with the SCORR1 option

[c] Percentage of patients in sample with treatment propensity score greater than .05 and less than .95 when only X_{1i}, X_{2i}, X_{3i}, X_{4i} factors are specified in the propensity score equation

positive properties of CFAs extend to real-world clinical applications in which patient factors affecting treatment effectiveness also influence treatment choice. In many real-world clinical scenarios it is plausible and likely that observed treatment choices reflect unmeasured patient factors related to expected treatment effectiveness for each patient – a condition defined in econometric literature as *essential heterogeneity* [38, 39, 43, 48–50, 53]. This paper used simulations that varied only by the relationship between treatment effectiveness and treatment choice to assess the impact of essential heterogeneity on the ability of CFAs to estimate patient-specific treatment effects. The causal forest algorithm within the generalized random forests application CFA-GRF has been singled out as most appropriate CFA estimate patient-specific treatment effects and was used here [98]. To tease out the impacts of essential heterogeneity, CFA-GRF estimates were evaluated in settings in which all patient factors associated with treatment effect heterogeneity were fully observed by the researcher and in settings in which the patient factors associated with treatment effect heterogeneity were not fully observed by the researcher.

We replicated the positive properties of CFA-GRF in simulation scenarios under ignorability. CFA-GRF yielded average population-wide estimates and average estimates by patient subsets based on treatment choice under ignorability that were closely aligned with their true values whether heterogeneity was fully or partially observed within the algorithm. As a result, if researchers can make a strong conceptual case a priori that treatment effectiveness is unrelated to treatment choice, they can be confident that CFA-GRF can yield appropriate treatment effect estimates across a population of patients. In simulation scenarios in which decision-makers use patient factors associated with treatment effectiveness in making treatment decisions [38, 39, 43, 48–50, 53], the ability of CFA-GRF to identify patient-specific treatment effects varied with the influence that treatment effectiveness had on treatment choice and whether the full range of patient factors associated with treatment effect heterogeneity were observed and specified in the algorithm. When all patient factors affecting treatment effect heterogeneity were fully specified, CFA-GRF produced treatment effect estimates that reflected true treatment effects across each population subset when the influence of treatment effectiveness on treatment choice was low. As this influence increased, however, treatment effect estimates showed increasingly negative bias for treated patients and positive bias for untreated patients. A substantial portion of this bias is likely attributable to nonoverlapping patients becoming a higher percentage of patients as the influence of treatment effectiveness on treatment choice increases. Under partially observed heterogeneity, all patients overlapped

in all simulations. CFA-GRF produced estimates that closely reflected the true treatment effect values for treated patients across all levels of influence of treatment effectiveness on treatment choice. In contrast, CFA-GRF estimates for untreated patients were biased high, with the extent of this bias increasing with the level of influence that treatment effectiveness had on treatment choice.

As a result, CFA-GRF estimates of patient-specific treatment effects using observational data must be assessed through the prism of the assumed reasons why patients with similar measured factors in a real-world setting were observed making different treatment choices. This requires researchers to explicitly develop conceptual frameworks of treatment choice to support these assumptions a priori to ensure proper interpretation of treatment effect estimates *ex post*. The call for treatment choice conceptual frameworks to guide treatment effectiveness research using observational data has long been stated in economics [44, 48, 49, 108–110], and the importance of these frameworks is now being more widely appreciated [21, 111, 112]. A conceptual framework of treatment choice should describe the factors thought to influence treatment choice, the relationship of these factors to treatment effectiveness and whether these factors are measured within the available data. Given the study findings, it would be important for researchers to qualify patient-specific estimates from CFA-GRF in clinical scenarios in which essential heterogeneity likely exists. In these scenarios researchers should state that patient-specific estimates from CFA-GRF are likely biased high for the average patient with a given combination measured patient factors and are best aligned to those patients a provider is more likely to treat.

This study is limited by its use of only using one of the several CFAs available to produce patient-specific evidence using observational data. While the CFA-GRF was singled out as most appropriate for estimating patient-specific treatment effects [98], it is possible that other CFAs are available that can incorporate and correct for the conditions associated with treatment choice when making treatment effect estimates. To this end, the simulated datasets produced here are available from the authors for use by other CFA developers to assess the impact on treatment effect estimates of the influence of treatment effect heterogeneity on treatment choice. In addition, the simulation approach in this paper is reported fully, is straightforward to reproduce, and is easy to modify, so researchers can assess the robustness of our results to parameter changes.

Conclusion

The acknowledged breadth of *treatment effect heterogeneity* across patients heightens the need to find empirical approaches to find patient-specific treatment effect evidence [4–10]. Causal forest algorithms (CFAs) have been proposed to analyze the treatment variation found within large observational databases to develop patient-specific evidence [23–33]. The simulation results in this paper show that the patient-specific estimates produced by a CFA are sensitive to the reasons why patients with the same set of measured factors were observed to make different treatment choices. It is likely in many real-world clinical scenarios that decision-makers are cognizant of how patient factors affect treatment effectiveness and use this information in making treatment decisions [38, 39, 43, 48–50, 53]. And many real-world decision makers may know more about the list of patient factors affecting treatment effectiveness than the researchers who collect measures for research [22, 113, 114]. As a result, it is foundational that researchers using CFAs to estimate patient-specific evidence using observational data build conceptual frameworks of treatment choice prior to estimation to guide estimate interpretation ex post.

Abbreviations

CFA	Causal forest algorithm
ATT	Average treatment effect on the treated
DAG	Directed acyclic graph
CART	Classification and regression tree
CFA-GRF	Causal forest algorithm - generalized random forest application

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02187-5.

Supplementary Material 1.

Acknowledgements

The authors acknowledge the support of the University of South Carolina Big Data Health Science Center and the Center for Effectiveness Research in Orthopaedics.

Authors' contributions

JMB created the simulation scenarios in the paper with conceptual and programming guidance from CGC, BKC, SF, and NH. JMB wrote the first draft of the manuscript with BKC, CGC, SF, and NH providing key insightful editorial changes in focus and direction.

Funding

This project was generously funded by a grant from the University of South Carolina Big Data Health Science Center and focused funding from the Center for Effectiveness Research in Orthopaedics.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval approval and consent to participate
This study uses simulated data with no human interaction. As such, this study was designated "exempt" by the University of South Carolina Institutional Review Board under Category 4 of 45 CFR 46.101(2)(b). All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable. Because this study had no human interaction, informed consent was deemed unnecessary according to national regulations by the University of South Carolina Institutional Review Board.

Competing interests
The authors declare no competing interests.

Received: 28 November 2023 Accepted: 21 February 2024
Published online: 13 March 2024

References

1. Patient Centered Outcomes Research Institute. Our Programs. https://www.pcori.org/about-us/our-programs. Published 2017. Accessed 20 Mar 2019.
2. Selby JV, Whitlock EP, Sherman KS, Slutsky JR. The Role of Comparative Effectiveness Research. In: Gallin JL, Ognibene FP, Johnson LL, editors. Principles and Practice of Clinical Research. 4th ed. London, UK: Elsevier; 2018. p. 269–92.
3. Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. Jama-J Am Med Assoc. 2012;307(15):1583–4.
4. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82(4):661–87.
5. Lohr KN, Eleazer K, Mauskopf J. Health policy issues and applications for evidence-medicine and clinical practice guidelines. Health Policy. 1998;46:1–19.
6. Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176–86.
7. Starfield B. Threads and yarns: weaving the tapestry of comorbidity. Ann Fam Med. 2006;4(2):101–3.
8. Steinberg EP, Luce BR. Evidence based? Caveat emptor! Health Affair. 2005;24(1):80–92.
9. Upshur REG. Looking for rules in a world of exceptions. Perspect Biol Med. 2005;48(4):477–89.
10. Dubois RW. From methods to policy: a "one-size-fits-all" policy ignores patient heterogeneity. J Comp Eff Res. 2012;1(2):119–20.
11. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172(1):35–45.
12. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med. 2018;210:2–21.
13. Concato J, Horwitz RI. Randomized trials and evidence in medicine: A commentary on deaton and cartwright. Soc Sci Med. 2018;210:32–6.
14. Rekkas A, Paulus JK, Raman G, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. BMC Med Res Methodol. 2020;20(1):264.
15. Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Publ Health. 2012;33:425–45.
16. Kowalski CJ, Mrdjenovich AJ. Comparative effectiveness research: decision-based evidence. Perspect Biol Med. 2014;57(2):224–48.
17. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int J Epidemiol. 2016;45(6):2184–93.
18. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;363:k4245.
19. Kent DM, van Klaveren D, Paulus JK, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement: explanation and elaboration. Ann Intern Med. 2020;172(1):W1–25.
20. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2020;41:21–36.
21. Crown WH. Real-world evidence, causal inference, and machine learning. Value Health. 2019;22(5):587–92.
22. Dekkers OM, Mulder JM. When will individuals meet their personalized probabilities? A philosophical note on risk prediction. Eur J Epidemiol. 2020;35(12):1115–21.
23. Athey S. Beyond prediction: using big data for policy problems. Science. 2017;355(6324):483–5.
24. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47(2):1148–78.
25. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci. 2016;113(27):7353–60.
26. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113(523):1228–42.
27. Bargagli-Stoffi FJ, De-Witte K, Gnecco G. Heterogeneous causal effects with imperfect compliance: a novel Bayesian machine learning approach. arXiv preprint arXiv:190512707. 2019.
28. Stoffi FJB, Gnecco G. Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. Paper presented at: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)2018.
29. Johnson M, Cao J, Kang H. Detecting heterogeneous treatment effect with instrumental variables. arXiv preprint arXiv:190803652. 2019.
30. Bargagli-Stoffi FJ, Gnecco G. Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. Int J Data Sci Analytics. 2020;9(3):315–37.
31. Wang G, Li J, Hopp W, J. An Instrumental Variable Forest Approach for Detecting Heterogeneous Treatment Effects in Observational Studies. Management Science. 2021;https://doi.org/10.1287/mnsc.2021.4084.
32. Dusseldorp E, Doove L, Mechelen I. Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. Behav Res Methods. 2016;48(2):650–63.
33. Su XG, Tsai CL, Wang HS, Nickerson DM, Li BG. Subgroup analysis via recursive partitioning. J Mach Learn Res. 2009;10:141–58.
34. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. P Natl Acad Sci USA. 2016;113(27):7353–60.
35. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Stat Med. 2018;37(23):3309–24.
36. Hahn PR, Dorie V, Murray JS. Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017. 2019:arXiv:1905.09515. https://doi.org/10.48550/arXiv.1905.09515. Accessed 1 May 2019.
37. Jawadekar N, Kezios K, Odden MC, et al. Practical guide to honest causal forests for identifying heterogeneous treatment effects. Am J Epidemiol. 2023;192(7):1155–65.
38. Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ. 2007;16(11):1133–57.
39. Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. Rev Econ Stat. 2006;88(3):389–432.
40. Basu A. Estimating Decision-Relevant Comparative Effects Using Instrumental Variables. Stat Biosci. 2011;3(1):6–27.
41. Ravallion M. On the implications of essential heterogeneity for estimating causal impacts using social experiments. J Econ Methods. 2015;4(1):145–51.
42. Heckman J, Pinto R. The econometric model for causal policy analysis. Annu Rev Econom. 2022;14(1):893–923.
43. Brooks JM, Chapman CG, Schroeder MC. Understanding treatment effect estimates when treatment effects are heterogeneous for more than one outcome. Appl Health Econ Health Policy. 2018;16(3):381–93.
44. Heckman JJ. Econometric causality. Int Stat Rev. 2008;76(1):1–27.
45. Heckman JJ, Vytlacil E. Structural equations, treatment effects, and econometric policy evaluation. Econometrica. 2005;73(3):669–738.

46. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceed National Acad Sci United States. 1999;96(8):4730–4.
47. Basu A. Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. Stata J. 2015;15(2):397–410.
48. Brooks JM, Fang G. Interpreting treatment-effect estimates with heterogeneity and choice: simulation model results. Clin Ther. 2009;31(4):902–19.
49. Garrido MM, Dowd B, Hebert PL, Maciejewski ML. Understanding treatment effect terminology in pain and symptom management research. J Pain Symptom Manage. 2016;52(3):446–52.
50. Smith J, Sweetman A. Viewpoint: estimating the causal effects of policies and programs. Can J Econ. 2016;49(3):871–905.
51. Heckman JJ. Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. J Polit Econ. 2001;109(4):673–748.
52. Angrist JD, Pischke J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, New Jersey: Princeton University Press; 2009.
53. Chapman CG, Brooks JM. Treatment effect estimation using nonlinear two-stage instrumental variable estimators: another cautionary note. Health Serv Res. 2016;51(6):2375–94.
54. Brooks JM, Chrischilles EA. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. Med Care. 2007;45(10 Supl 2):123–30.
55. Angrist JD, Ferandez-Val I. ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework. In: Acemoglu D, Arellano M, Dekel E, eds. Advances in Economics and Econometrics, Vol Iii: Econometrics.2013:401–433.
56. Angrist JD. Treatment effect heterogeneity in theory and practice. Econ J. 2004;114:C52–83.
57. Heckman JJ, Robb R. Alternative Methods for Evaluating the Impact of Interventions. In: Heckman JJ, Singer B, editors. Longitudinal Analysis of Labor Market Data. New York: Cambridge University Press; 1985. p. 156–245.
58. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica. 1994;62(2):467–75.
59. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444–55.
60. Angrist JD. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. J Business Econ Statistics. 2001;19(1):2–16.
61. Moler-Zapata S, Grieve R, Basu A, O'Neill S. How does a local instrumental variable method perform across settings with instruments of differing strengths? A simulation study and an evaluation of emergency surgery. Health Econ. 2023;32(9):2113–26.
62. Brooks JM, Chapman CG, Cozad MJ. The identification process using choice theory is needed to match design with objectives in CER. Med Care. 2017;55(2):91–3.
63. Cozad MJ, Chapman CG, Brooks JM. Specifying a conceptual treatment choice relationship before analysis is necessary for comparative effectiveness research. Med Care. 2016;55(2):94–6.
64. Heckman JJ. The scientific model of causality. Sociol Methodol. 2005;35:1–97.
65. Angrist JD. Treatment effect heterogeneity in theory and practice. Econ J. 2003;114:1–30.
66. Manski CF. [Choices as an alternative to control in observational studies]: comment. Stat Sci. 1999;14(3):279–81.
67. Harris KM, Remler DK. Who is the marginal patient? understanding instrumental variables estimates of treatment effects. Health Serv Res. 1998;33(5):1337–60.
68. Heckman JJ, Robb R. Alternative methods for evaluating the impact of interventions - an overview. J Econ. 1985;30(1–2):239–67.
69. Blundell R, Costa DM. Evaluation methods for non-experimental data. Fisc Stud. 2000;21(4):427–68.
70. Smith J. Treatment effect heterogeneity. Eval Rev. 2022;46(5):652–77.
71. Brooks JM, Chrischilles EA. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. Med Care. 2007;45(10):S123–30.
72. Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
73. Jayakumar P, Teunis T, Williams M, Lamb SE, Ring D, Gwilym S. Factors associated with the magnitude of limitations during recovery from a fracture of the proximal humerus predictors of limitations after proximal humerus fracture. Bone Joint J. 2019;101(6):715–23.
74. Otlans PT, Szukics PF, Bryan ST, Tjoumakaris FP, Freedman KB. Current concepts review resilience in the orthopaedic patient. J Bone Joint Surg-Am. 2021;103(6):549–59.
75. Ezeamama AE, Elkins J, Simpson C, Smith SL, Allegra JC, Miles TP. Indicators of resilience and healthcare outcomes: findings from the 2010 health and retirement survey. Qual Life Res. 2016;25(4):1007–15.
76. Floyd SB, Walker JT, Smith JT, et al. ICD-10 diagnosis codes in electronic health records do not adequately capture fracture complexity for proximal humerus fractures. J Shoulder Elbow Surg. 2023;33(2):417–24.
77. Floyd SB, Thigpen C, Kissenberth M, Brooks JM. Association of surgical treatment with adverse events and mortality among medicare beneficiaries with proximal humerus fracture. JAMA Netw Open. 2020;3(1):e1918663.
78. Brooks JM, Chapman CG, Floyd SB, Chen BK, Thigpen CA, Kissenberth M. Assessing the ability of an instrumental variable causal forest algorithm to personalize treatment evidence using observational data: the case of early surgery for shoulder fracture. BMC Med Res Methodol. 2022;22(1):190.
79. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care. 2012;50(3):217–26.
80. Landes SJ, McBain SA, Curran GM. An introduction to effectiveness-implementation hybrid designs. Psychiatry Res. 2019;280:112513.
81. Curran GM, Landes SJ, McBain SA, et al. Reflections on 10 years of effectiveness-implementation hybrid studies. Front Health Serv. 2022;2:1053496.
82. Wolfenden L, Williams CM, Wiggers J, Nathan N, Yoong SL. Improving the translation of health promotion interventions using effectiveness–implementation hybrid designs in program evaluations. Health Promot J Austr. 2016;27(3):204–7.
83. Bernet AC, Willens DE, Bauer MS. Effectiveness-implementation hybrid designs: implications for quality improvement science. Implement Sci. 2013;8(1):S2.
84. Ullman AJ, Beidas RS, Bonafide CP. Methodological progress note: Hybrid effectiveness-implementation clinical trials. J Hosp Med. 2022;17(11):912–6.
85. Liang YY, Ehler BR, Hollenbeak CS, Turner BJ. Behavioral support intervention for uncontrolled hypertension a Complier Average Causal Effect (CACE) Analysis. Med Care. 2015;53(2):E9–15.
86. Peugh JL, Strotman D, McGrady M, Rausch J, Kashikar-Zuck S. Beyond intent to treat (ITT): a complier average causal effect (CACE) estimation primer. J School Psychol. 2017;60:7–24.
87. Knox CR, Lall R, Hansen Z, Lamb SE. Treatment compliance and effectiveness of a cognitive behavioural intervention for low back pain: a complier average causal effect approach to the BeST data set. Bmc Musculoskeletal Dis. 2014;15:1–1.
88. Berg JK, Bradshaw CP, Jo B, Ialongo NS. Using Complier average causal effect estimation to determine the impacts of the good behavior game preventive intervention on teacher implementers. Adm Policy Ment Health. 2017;44(4):558–71.
89. Gruber JS, Arnold BF, Reygadas F, Hubbard AE, Colford JM Jr. Estimation of treatment efficacy with complier average causal effects (CACE) in a randomized stepped wedge trial. Am J Epidemiol. 2014;179(9):1134–42.
90. Connell AM. Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. Am J Drug Alcohol Abuse. 2009;35(4):253–9.
91. Ashworth E, Panayiotou M, Humphrey N, Hennessey A. Game on-complier average causal effect estimation reveals sleeper effects on academic attainment in a randomized trial of the good behavior game. Prev Sci. 2020;21(2):222–33.
92. Panayiotou M, Humphrey N, Hennessey A. implementation matters: using complier average causal effect estimation to determine the impact of the promoting alternative thinking strategies (PATHS) curriculum on children's quality of life. J Educ Psychol. 2020;112(2):236–53.
93. Carmody T, Greer TL, Walker R, Rethorst CD, Trivedi MH. A complier average causal effect analysis of the stimulant reduction intervention using dosed exercise study. Cont Clin Trial Comm. 2018;10:1–8.

94. Huang S, Cordova D, Estrada Y, Brincks AM, Asfour LS, Prado G. An application of the complier average causal effect analysis to examine the effects of a family intervention in reducing illicit drug use among high-risk hispanic adolescents. Fam Process. 2014;53(2):336–47.
95. Cowan JM. School choice as a latent variable: Estimating the "complier average causal effect" of vouchers in Charlotte. Policy Stud J. 2008;36(2):301–15.
96. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
97. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression Trees. CRC Press; 1984.
98. McConnell KJ, Lindner S. Estimating treatment effects with machine learning. Health Serv Res. 2019;54(6):1273–82.
99. Roy AD. Some thoughts on the distribution of earnings. Oxford Econ Pap. 1951;3(2):135–46.
100. Weinberg CR. Can DAGs clarify effect modification? Epidemiology. 2007;18(5):569–72.
101. Attia J, Holliday E, Oldmeadow C. A proposal for capturing interaction and effect modification using DAGs. Int J Epidemiol. 2022;51(4):1047–53.
102. Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399–424.
103. Walker AM, Patrick AR, Lauer MS, et al. A tool for assessing the feasibility of comparative effectiveness research. Comparative Effect Res. 2013;3:11–20.
104. Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. Am J Epidemiol. 2010;172(7):843–54.
105. Sturmer T, Webster-Clark M, Lund JL, et al. Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: a simulation study. Am J Epidemiol. 2021;190(8):1659–70.
106. Tibshirani J, Athey S, Sverdrup E, Wager S. instrumental_forest: Instrumental Forest. https://rdrr.io/cran/grf/man/instrumental_forest.html. Published 2021. Accessed 15 May 2021.
107. Sadique Z, Grieve R, Diaz-Ordaz K, Mouncey P, Lamontagne F, O'Neill S. A machine-learning approach for estimating subgroup- and individual-level treatment effects: an illustration using the 65 trial. Med Decis Making. 2022;42(7):923–36.
108. Cozad MJ, Chapman CG, Brooks JM. Specifying a conceptual treatment choice relationship before analysis is necessary for comparative effectiveness research. Med Care. 2017;55(2):94–6.
109. Lewbel A. The identification zoo: meanings of identification in econometrics. J Econ Lit. 2019;57(4):835–903.
110. Heckman JJ. Building bridges between structural and program evaluation approaches to evaluating policy. J Econ Lit. 2010;48(2):356–98.
111. Ho M, van der Laan M, Lee H, et al. The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. Statistics Biopharmaceut Res. 2021;15:1–14.
112. VanderWeele TJ, Mathur MB. Commentary: developing best-practice guidelines for the reporting of E-values. Int J Epidemiol. 2020;49(5):1495–7.
113. Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2018;100:22–31.
114. Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health. 2020;2(12):e677–80.

Publisher's Note