European Journal of Epidemiology (2023) 38:123–133
https://doi.org/10.1007/s10654-022-00901-5

ESSAY

# Regression-based estimation of heterogeneous treatment effects when extending inferences from a randomized trial to a target population

Sarah E. Robertson[1,2] · Jon A. Steingrimsson[3] · Issa J. Dahabreh[1,2,4]

Received: 4 October 2021 / Accepted: 11 July 2022 / Published online: 10 January 2023


## Abstract

Most work on extending (generalizing or transporting) inferences from a randomized trial to a target population has focused on estimating average treatment effects (i.e., averaged over the target population's covariate distribution). Yet, in the presence of strong effect modification by baseline covariates, the average treatment effect in the target population may be less relevant for guiding treatment decisions. Instead, the conditional average treatment effect (CATE) as a function of key effect modifiers may be a more useful estimand. Recent work on estimating target population CATEs using baseline covariate, treatment, and outcome data from the trial and covariate data from the target population only allows for the examination of heterogeneity over distinct subgroups. We describe flexible pseudo-outcome regression modeling methods for estimating target population CATEs conditional on discrete or continuous baseline covariates when the trial is embedded in a sample from the target population (i.e., in nested trial designs). We construct pointwise confidence intervals for the CATE at a specific value of the effect modifiers and uniform confidence bands for the CATE function. Last, we illustrate the methods using data from the Coronary Artery Surgery Study (CASS) to estimate CATEs given history of myocardial infarction and baseline ejection fraction value in the target population of all trial-eligible patients with stable ischemic heart disease.

**Keywords** Heterogeneity of treatment effect · Transportability · Generalizability · Conditional average treatment effect · Epidemiologic methods

## Abbreviations

| | |
|---|---|
| CASS | Coronary Artery Surgery Study |
| CATE | Conditional average treatment effect |
| MI | Myocardial infarction |

## Introduction

When treatment effect modifiers have a different distribution among participants in a randomized trial compared to the target population of substantive interest, the average treatment effect estimate from the trial is not directly applicable to the target population. A growing literature describes methods for extending — transporting or generalizing [1, 2] — inferences for the average treatment effect from the trial to the target population [3–7]. These methods critically depend on adjusting for a large number of covariates to ensure that the trial and target population are conditionally exchangeable, allowing estimation of the target population average treatment effect.

Yet, the target population average treatment effect may not be sufficient for guiding treatment or policy decisions in the presence of strong effect modification [8], especially when a small set of strong effect modifiers can be identified on the basis of background knowledge. In such cases, the target population conditional average treatment effect (CATE) as a function of these strong effect modifiers may be a more useful estimand [9]. For example, investigators may

✉ Issa J. Dahabreh
idahabreh@hsph.harvard.edu

1 CAUSALab, Harvard T.H. Chan School of Public Health, Boston, MA, USA

2 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

3 Department of Biostatistics, Brown University School of Public Health, Providence, RI, USA

4 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

be able to identify a few "causal" effect modifiers of primary interest and choose to estimate the target population CATE given these causal effect modifiers to guide clinical or policy decisions, while also recognizing that they need to adjust for many additional candidate effect modifiers, including "surrogate" ones, to render the trial and target populations exchangeable [10, 11].

Recent work [12, 13] has described methods for estimating target population CATEs over distinct subgroups formed by a few key discrete (or discretized) covariates (i.e., "subgroup-specific average treatment effects"), when extending inferences from the trial to the target population. For example, a g-formula approach (outcome model-based) involves estimating an outcome model for each treatment group in the trial, conditional on the large number of covariates needed for exchangeability, obtaining predictions under each model for the individuals in the target population, and averaging the difference between the predictions under each model within levels of the subgroup variable to estimate CATEs in the target population (see [13] for details). This g-formula approach and related weighting and augmented (doubly robust) weighting methods [13], however, cannot be used to estimate the target population CATE over continuous covariates or multiple discrete covariates, because the number of observations at each covariate level is not adequate for estimation, even in large data sets [11, 14].

In this paper, we describe methods for estimating the target population CATE given a small set of continuous or discrete effect modifiers. Specifically, we build on recent advances in estimating CATEs in observational studies [15–21] and efficient and robust methods for generalizability and transportability analyses [6, 7], to propose a flexible two-step regression procedure for estimating the target population CATEs when a trial is embedded in a sample from the target population (i.e., in nested trial designs [22]). In the first step of the procedure, a pseudo-outcome is formed using models for the conditional probability of trial participation and the conditional expectation of the outcome in the trial. In the second step, the pseudo-outcome is regressed on the key effect modifiers. The procedure can support valid inference about the CATE when using data-adaptive (e.g., machine learning) approaches to estimate the probability of trial participation and the expectation of the outcome (in the first step), and allows flexible modeling of the treatment effect as a function of the effect modifiers (in the second step). Thus, the procedure facilitates the examination of heterogeneity over a low-dimensional set of key effect modifiers, while allowing adjustment for a potentially high-dimensional set of covariates that is sufficient to render the trial and target population exchangeable (i.e., adjust for selective participation). We show how to construct pointwise confidence intervals for the CATE at a specific value of the key effect modifiers and uniform confidence bands for the CATE function. We illustrate the proposed methods using data from the Coronary Artery Surgery Study (CASS) to estimate CATEs given history of myocardial infarction and baseline ejection fraction value in the target population of trial-eligible patients with stable ischemic heart disease.

## Study design, data, and causal estimands

### Study design

We consider a nested trial design [22], where the trial is embedded within a cohort sampled from the target population of substantive interest. The nesting can be achieved by designing a prospective cohort study of individuals from the target population and inviting some of the cohort members to participate in the trial, while collecting information on baseline covariates on all cohort members, including those who are not invited or do not agree to participate in the trial. Nesting can also be achieved by retrospectively linking records from a completed trial with records from a cohort that is sampled from the target population. Regardless of how nesting is achieved, we assume that the cohort in which the trial is nested can be viewed as a simple random sample from the target population [23]. Nested trial designs can be used for generalizability analyses, when the target population represented by the cohort meets the trial eligibility criteria, as well as for transportability analyses, when the target population represented by the cohort is broader than the population defined by the trial's eligibility criteria (see [1, 2] for details regarding the definitions of the terms generalizability and transportability that we use in this paper; these terms are not used consistently in the literature, e.g., reference [4] suggests different definitions).

### Data

In the nested trial design, data on a vector of baseline covariates, $X$, are available from all individuals in the cohort, regardless of participation in the trial. Data on the assigned treatment, $A$, and the outcome, $Y$, need only be available among trial participants. We use $S$ as the indicator for trial participation ($S = 1$ for randomized individuals; $S = 0$ for non-randomized individuals). Thus, the observed data are realizations of independent random draws of the tuple $O_i = (X_i, S_i, S_iA_i, S_iY_i)$, for $i = 1, \ldots, n$, where $n$ is the total number of individuals in the cohort (both randomized and non-randomized). For simplicity, we assume throughout that the treatment is binary; extensions to address multi-valued treatments are similar to the binary case but require a slightly different set of identification conditions and modifications

to the estimation methods (e.g., to use a "generalized" propensity score approach [15] to estimate the probability of treatment; see [21] for related results using multi-valued treatments in observational studies). We also assume that the outcome is measured at the end of the study (binary, continuous, or count; we do not consider failure-time outcomes in this paper). Throughout, italic upper-case letters denote random variables and corresponding lower-case letters denote realizations. We use $f(\cdot)$ to generically denote densities.

### Causal estimands

Let $Y^a$ denote the potential (counterfactual) outcome under each treatment $a \in \{0, 1\}$, that is, the outcome that would be observed under intervention to set treatment to $a$ [24–26]. Furthermore, let $\widetilde{X}$ denote a vector that contains a small subset of the covariates in $X$ that, on the basis of mechanistic understanding and prior empirical evidence, are *a priori* considered as the "key" effect modifiers under consideration. We are interested in the target population CATE given $\widetilde{X} = \widetilde{x}$,

$$\mathrm{E}\left[Y^1 - Y^0 \middle| \widetilde{X} = \widetilde{x}\right] = \mathrm{E}\left[Y^1 \middle| \widetilde{X} = \widetilde{x}\right] - \mathrm{E}\left[Y^0 \middle| \widetilde{X} = \widetilde{x}\right].$$

The CATE at some specific value of the key effect modifier, $\widetilde{x}$, is sometimes referred to as the "group-average treatment effect" [17].

In generalizability and transportability analyses, the covariate vector $X$ may be high-dimensional because investigators collect information on multiple covariates in the hope of rendering the trial and the target population exchangeable (in the sense formalized in the next section). In contrast, the vector of key effect modifiers $\widetilde{X}$ is typically low-dimensional, containing only the small subset of baseline covariates that are deemed to be key effect modifiers of interest. For instance, in our illustrative example, the CASS investigators identified history of myocardial infarction and abnormal left ventricular function (defined as ejection fraction <50%) as key effect modifiers and examined them in subgroup analyses using data from trial participants [27, 28]. Thus, we may want to examine the association between the treatment effect in the target population and history of myocardial infarction and ejection fraction. Such examination, however, is likely to first require conditioning on many additional covariates to render the trial participants exchangeable with the target population.

## Identification

### Identifiability conditions

The following conditions, which are sufficient for identifying the average treatment effect in the target population [6, 29], are also sufficient for identifying the CATE in the target population:

(1) *Consistency of potential outcomes:* if $A_i = a$, then $Y_i = Y_i^a$, for each $a \in \{0, 1\}$ and for every individual $i$ in the target population.
(2) *Mean exchangeability over A in the trial:* for each $a \in \{0, 1\}$ and every $x$ with positive density in the trial $f(x, S = 1) \neq 0$, $\mathrm{E}[Y^a | X = x, S = 1, A = a] = \mathrm{E}[Y^a | X = x, S = 1]$.
(3) *Positivity of the treatment probability in the trial:* $\Pr[A = a | X = x, S = 1] > 0$ for each $a \in \{0, 1\}$ and every $x$ with positive density in the trial $f(x, S = 1) \neq 0$.
(4) *Exchangeability in effect measure over S:* $\mathrm{E}[Y^1 - Y^0 | X = x, S = 1] = \mathrm{E}[Y^1 - Y^0 | X = x]$, for every $x$ with positive density in the target population $f(x) \neq 0$.
(5) *Positivity of trial participation:* $\Pr[S = 1 | X = x] > 0$, for every $x$ with positive density in the target population $f(x) \neq 0$.

The consistency condition over all individuals in the target population implicitly requires the absence of "hidden" versions of treatment [30–32], trial engagement effects [2, 29], and interference [30, 33]. These conditions are largely untestable and need to be considered on the basis of substantive knowledge. The conditions of mean exchangeability and positivity of the treatment probability are expected to hold in (marginally or conditionally) randomized trials comparing well-defined interventions [11]. The condition of "exchangeability in effect measure over $S$" reflects an untestable assumption of lack of effect measure modification by trial participation, conditional on baseline covariates, and needs to be examined in light of substantive knowledge and subjected to sensitivity analyses [34]. Stronger assumptions of exchangeability in expectation or in distribution between the trial and target population allow identification of other estimands that are not identifiable under exchangeability in measure (e.g., potential outcome means) [3, 6, 35]. Last, positivity of trial participation is in principle testable, but its assessment can be challenging when $X$ is high-dimensional [36].

To focus on issues related to extending inferences from trials, we assume that there are no missing data, no losses to follow-up, and complete adherence to treatment. The

methods we describe can be extended to address these complications, under additional assumptions [29] and provided additional data are collected [37]. Furthermore, our results also apply to generalizing or transporting inferences from an observational study nested in a broader cohort, provided we are willing to assume that conditions (2) and (3) hold in the observational study (i.e., no unmeasured baseline confounding and positivity of treatment within levels of the measured confounders) [38].

### Identification of CATEs

In Web Appendix 1, we show that, under conditions (1) through (5), the target population CATE given $\widetilde{X} = \widetilde{x}$, that is, $\mathrm{E}\left[Y^1 - Y^0 \middle| \widetilde{X} = \widetilde{x}\right]$, is identified by

$$\delta(\widetilde{x}) \equiv \mathrm{E}\left[\phi(O) \middle| \widetilde{X} = \widetilde{x}\right]; \tag{1}$$

where the pseudo-outcome $\phi(O)$ is defined as

$$\phi(O) = \phi_1(O) - \phi_0(O),$$

with

$$\phi_a(O) = \frac{SI(A=a)}{p(X)e_a(X)}\left\{Y - g_a(X)\right\} + g_a(X), \text{ for } a = 0, 1;$$

$p(X) = \Pr[S = 1|X]$ is the conditional probability of trial participation given covariates; $e_a(X) = \Pr[A = a|X, S = 1]$ is the conditional probability of treatment $a$ in the trial given covariates; and $g_a(X) = \mathrm{E}[Y|X, S = 1, A = a]$ is the conditional expectation of the outcome in the trial given covariates, among individuals assigned to treatment $a \in \{0, 1\}$. We refer to the functions $p(X)$, $e_a(X)$, and $g_a(X)$ as "nuisance functions" because they are useful in identifying and estimating CATEs, but, in our setup, are not of scientific interest *per se*. We refer to $\phi(O)$ generically as a "pseudo-outcome" because it is a constructed variable (using the models for the probability of participation, the probability of treatment, and the expectation of the outcome) that is used as a "response" (i.e., "left-hand-side" variable) in the regression of equation (1). Because $\phi(O)$ involves the observed data $O$ and nuisance functions that are identifiable from the observed data under the nested trial design [22], we conclude that CATEs given $\widetilde{X}$ are identifiable. Of note, $\phi(O)$ is the (uncentered) influence function [39] of the functional that identifies the target population average treatment effect under a nonparametric model for the observed data that obeys conditions (1) through (5); see reference [6] for details.

An inverse probability weighted pseudo-outcome is formed by setting the $g_1(X)$ and $g_0(X)$ terms to zero in the expression for $\phi(O)$ (see Web Appendix 1 and Web Appendix 2 for details). For the remainder of this paper, we do not consider this simpler pseudo-outcome because it does not allow for valid inference when using machine learning to estimate the nuisance functions [40].

## Estimation & inference

### Two-step estimation procedure

To extend causal inferences about CATEs given key effect modifiers from the trial to the target population, we propose a two-step procedure, similar to approaches for estimating CATEs in observational analyses with baseline confounding by measured variables [19, 20]. In the first step, we create the pseudo-outcome using models for the probability of trial participation, the expectation of the outcome, and (optionally) the probability of treatment in the trial, to account for differences between the trial and the target population by conditioning on a large set of effect modifiers. In the second step, we regress the pseudo-outcome on the key effect modifiers to estimate the CATE function.

**Step 1: Estimation of the nuisance functions to form the pseudo-outcome:** Forming the pseudo-outcome for each observation $i = 1, \ldots, n$ in the data follows the identification results for $\phi(O)$ by using estimators for the nuisance functions (denoted by "hats"):

$$\widehat{\phi}(O_i) = \widehat{\phi}_1(O_i) - \widehat{\phi}_0(O_i), \tag{2}$$

with

$$\widehat{\phi}_a(O_i) = \frac{S_i I(A_i = a)}{\widehat{p}(X_i)\widehat{e}_a(X_i)}\left\{Y_i - \widehat{g}_a(X_i)\right\} + \widehat{g}_a(X_i), \text{ for } a = 0, 1.$$

Of note, the average of $\widehat{\phi}(O_i)$ over the $n$ observations in the sample, that is, $n^{-1}\sum_{i=1}^{n}\widehat{\phi}(O_i)$ gives a "doubly robust" or "augmented inverse probability weighted" [41] estimator of the average treatment effect in the target population (to see this, compare $\widehat{\phi}_a(O_i)$ with the summand in equation (5) of reference [6]).

When calculating the pseudo-outcomes, there are several options for estimating the nuisance functions for the probability of participation, expectation of the outcome, and the probability of treatment, that is, $p(X)$, $g_a(X)$, and $e_a(X)$, respectively. The probability of treatment in the trial, $e_a(X)$, is typically known by design so its estimation is straightforward using simple parametric models (e.g., logistic regression) [42, 43]. Alternatively, the known probability of treatment can be used. In contrast, the functions $p(X)$ and $g_a(X)$ are unknown, involve conditioning on the high-dimensional baseline covariates that are necessary for exchangeability to hold between the trial and the target population, and may be

complex, perhaps including nonlinearities and interactions. In practice, parametric models are commonly used to estimate the probability of participation, $p(X)$, or the expectation of the outcome, $g_a(X)$. When parametric models are used, the procedure is "model doubly robust" [41], in the sense that the CATE function can be consistently estimated when at least one of the parametric models for participation or the outcome is correctly specified. Nevertheless, parametric models may poorly approximate both $p(X)$ and $g_a(X)$. At the same time, the high-dimension of $X$ precludes fully nonparametric estimation of these models [14] (e.g., non-smooth nonparametric (frequency) estimation of the expectation of $Y$ given $X$ is infeasible if $X$ is high-dimensional or has continuous components [44]). To make progress, investigators can instead use machine learning (data-adaptive) methods to reduce model misspecification and allow more flexible modeling of the nuisance functions.

The cost of using data-adaptive approaches is that they converge to the true underlying nuisance function at a slower rate than parametric models. Informally, for valid inference, a fast enough rate of convergence of the data-adaptive approaches to the underlying true function is important to ensure that the bias of the estimated CATE function is "small" relative to its standard error [45]. Without a fast enough rate of convergence for the nuisance functions, bias remains, resulting in an inconsistent estimator without optimal coverage. To avoid bias when using data-adaptive approaches, by combining models for the probability of participation and the expectation of the outcome in the construction of the pseudo-outcome, we can rely on estimators of the nuisance functions that converge at a "fast enough," even if slower than parametric, combined rate (i.e., the estimator of the pseudo-outcome in equation (2) has a "rate-robustness property" [46]). Several data-adaptive methods can have rates that are fast enough (e.g., the highly adaptive least absolute shrinkage and selection operator (HAL) [47] and generalized additive models (GAMs) [48, 49]). When using data-adaptive approaches to estimate the nuisance functions for the probability of participation and the expectation of the outcome, we assume that the chosen approaches are consistent for the true underlying functions.

Background knowledge about aspects of the data-generating process can be used to select approaches that produce good approximations of the nuisance functions. For example, if we expect the relationship between trial participation and covariates, or the outcome and covariates, to be highly nonlinear or involve statistical interactions among covariates, random forest methods [50] may be a good choice to estimate the nuisance functions. If we expect sparsity, the least absolute shrinkage and selection operator [51] or other sparsity-appropriate modeling approaches may be preferred.

Regardless of the estimation approach (i.e., whether using parametric or data-adaptive approaches), the estimated functions are used to calculate the pseudo-outcomes in equation (2). Importantly, the participation and outcome models should include the high-dimensional set of variables needed to satisfy condition (4) to make the trial and target population conditionally exchangeable.

**Step 2: Pseudo-outcome regression:** We fit a regression of the estimated pseudo-outcome on the key effect modifiers, $\widetilde{X}$, to estimate the target population CATE as a function of $\widetilde{x}$,

$$\widehat{\delta}(\widetilde{x}) = \widehat{\mathrm{E}}\left[\widehat{\phi}(O)\middle|\widetilde{X} = \widetilde{x}\right]. \tag{3}$$

We refer to this second step of the procedure as a "pseudo-outcome regression." To consistently estimate the target population CATE function, we need to correctly specify the regression model in the second step, and have consistent estimators of the nuisance functions in the first step. One approach is to use a parametric model (e.g., least squares regression) to model the relationship between the pseudo-outcome and the key effect modifiers $\widetilde{X}$, but correct model specification may be challenging if $\widetilde{X}$ contains continuous components that are not guaranteed to be linear and may have complex functional forms. To mitigate the risk of model misspecification, given that in our setup $\widetilde{X}$ is low dimensional, it will often be possible to use nonparametric regression to flexibly model the relationship between the pseudo-outcome and the key effect modifiers, $\widetilde{X}$. For example, when $\widetilde{X}$ contains discrete covariates, it is simple to split the data into subgroups defined by the different levels of the covariates and estimate the mean of the pseudo-outcome within each subgroup – a non-smooth nonparametric "regression" approach [44]. When $\widetilde{X}$ also contains continuous components, we can use smoothing nonparametric techniques such as series [19] or kernel (local linear) regression [20] methods. For example, if $\widetilde{X}$ consists of a single continuous covariate, we can use a series estimator in the second step by fitting an ordinary least squares regression of the pseudo-outcome on a flexible polynomial (alternatively, we can use splines or other basis functions) [19]. The goal is to approximate the CATE function with a flexible model that is easy to understand and graph.

## Inference

We now discuss how to obtain both pointwise confidence intervals and uniform confidence bands. Pointwise confidence intervals are appropriate for the estimated CATE at a specific value of the key effect modifiers. These intervals capture the uncertainty at a specific point (i.e., at a specific value of the covariates $\widetilde{X}$) but are too narrow (will have significant undercoverage) if inferences are drawn over multiple points, as is the case when examining the entire domain of the CATE function. If investigators are interested

in examining heterogeneity across the entire domain of the CATE function, uniform confidence bands reflect uncertainty over multiple points. The inference strategies we describe are appropriate when the CATE function in the second step of the two-step estimation procedure is estimated using least squares, series, or kernel regression, provided appropriate technical requirements are met (e.g., regularity conditions, undersmoothing in the kernel regression, etc.; see [19] for details regarding ordinary least squares or series regression, and [20] for kernel regression). Statistical inference when using other approaches in the second step of the two-step estimation procedure would require case-by-case examination.

**Pointwise inference:** Pointwise confidence intervals for the CATE at a specific $\widetilde{x}$ value can be obtained using standard approaches, under reasonable technical conditions [19]. Specifically, a $(1-\alpha)\%$ pointwise confidence interval at $\widetilde{x}$ is given by $(\widehat{\delta}(\widetilde{x}) \pm z_{1-\alpha/2} \times \widehat{\sigma}(\widetilde{x}))$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and $\widehat{\sigma}(\widetilde{x})$ is the estimated standard error of the CATE at $\widetilde{x}$, which can be obtained using the nonparametric bootstrap [52]. Alternatively, when using ordinary least squares or series regression in the second step of the CATE estimation procedure, the robust variance estimator (i.e., the Huber-White sandwich estimator) [53, 54], which is readily available in standard software packages, can be used to obtain $\widehat{\sigma}(\widetilde{x})$. When using the pseudo-outcomes defined in equation (2), it is not necessary to account for the uncertainty in the fitted nuisance models in the first step when estimating the robust variance; inference can be carried out as if the nuisance functions were known [, 19].

**Uniform inference:** Uniform confidence bands are needed to obtain valid coverage over the set of values of the key effect modifiers that we16 want to examine. Suppose, for concreteness, that we are using series regression in the second step of the CATE estimation procedure. Series regression involves an estimator of the CATE as a function of $\widetilde{x}$ with the form $\widehat{\delta}(\widetilde{x}) = m(\widetilde{x})\widehat{\beta}$, where $m(\widetilde{x})$ is a vector of series or sieve basis functions (e.g., polynomials, splines, or wavelets) of $\widetilde{x}$, and $\widehat{\beta}$ is the least squares estimator of the regression coefficients. We will evaluate $\widehat{\delta}(\widetilde{x})$ over a set of grid points $\mathcal{P}$, where $\mathcal{P}$ is a subset of the possible values of the effect modifiers $\widetilde{X}$. We work with the grid points in $\mathcal{P}$ instead of all possible values of $\widetilde{X}$ to allow for the possibility that some components of $\widetilde{X}$ are continuous. One option is to use as grid points all the unique values of $\widetilde{x}$ observed in the data; another is to choose grid points that capture the "interesting" values of $\widetilde{x}$. To obtain uniform inference over the set $\mathcal{P}$, following [19, 20], we use a weighted bootstrap procedure [55–57] (sometimes referred to as the wild or multiplier bootstrap). We describe the weighted bootstrap procedure in Web Appendix 3; additional considerations for data-adaptive approaches are discussed in Web Appendix 4.

**Using a two-step approach to model conditional potential outcome means and other CATE measures:** See Web Appendix 5 for how the two-step procedure can be modified to model conditional potential outcome means (e.g., conditional counterfactual risks) and treatment effect measures other than the mean/risk difference.

## Examining heterogeneity in cass

### CASS design and data

The Coronary Artery Surgery Study (CASS) [27, 58] compared coronary artery bypass grafting surgery plus medical therapy (hereafter, "surgery") versus medical therapy alone in a randomized trial that was nested within a cohort of trial-eligible patients with stable ischemic heart disease. Patients were enrolled from August 1975 to May 1979 and followed-up for death up to December 1996. The cohort consisted of 2099 trial-eligible patients of whom 780 participated in the trial. For the first 10 years of follow-up, there was no censoring among trial participants.

The original CASS analysis prespecified variables that the investigators believed to be important effect modifiers and risk factors for the outcome of mortality [27]. These variables included history of myocardial infarction and abnormal left ventricular function (defined as ejection fraction value less than 50%). One analysis of the trial participants in CASS at 10-years of follow-up [59] found no difference in survival probability between treatment groups in the overall trial sample, but found that patients with an ejection fraction less than 50% had significantly improved survival with surgery (surgery 79%, medical therapy 61%). No other subgroup-specific benefits were found in the trial [28, 59]. A re-analysis of both randomized and observational data from CASS found heterogeneity on the risk difference scale for mortality at 10-years of follow-up among subgroups defined by history of myocardial infarction and abnormal left ventricular function [60]. Furthermore, a meta-analysis of 7 early trials (including CASS) comparing surgery versus medical therapy found that abnormal left ventricular function was an important modifier for the effect of treatment on mean survival time and that patients with abnormal left ventricular function derived greater absolute benefit from surgery [61]. A more recent randomized trial reported that among patients with ischemic cardiomyopathy and low ejection fraction (<35%), surgery was more beneficial than medical therapy [62]. Thus, we decided to use the methods described above to explore whether history of myocardial infarction and ejection fraction (treated as a continuous

**Table 1** Baseline characteristics in CASS (August 1975 to December 1996). $S = 1$ indicates the randomized group; $S = 0$ indicates the non-randomized group; $A = 1$ indicates surgery; $A = 0$ indicates medical therapy

| | $S = 1, A = 1$ | $S = 1, A = 0$ | $S = 1$ | $S = 0$ |
|---|---|---|---|---|
| Number of patients | 368 | 363 | 731 | 955 |
| Age | 51.42 (7.24) | 50.92 (7.41) | 51.17 (7.32) | 50.89 (7.73) |
| History of angina | 285 (77.4) | 282 (77.7) | 567 (77.6) | 760 (79.6) |
| Taken beta-blocker regularly | 163 (44.3) | 152 (41.9) | 315 (43.1) | 508 (53.2) |
| Taken diuretic regularly | 63 (17.1) | 50 (13.8) | 113 (15.5) | 145 (15.2) |
| Ejection fraction | 60.86 (13.04) | 59.83 (12.78) | 60.35 (12.91) | 60.16 (12.25) |
| Employed full-time | 264 (71.7) | 233 (64.2) | 497 (68.0) | 632 (66.2) |
| Type of job | | | | |
| High physical labor job | 151 (41.0) | 142 (39.1) | 293 (40.1) | 340 (35.6) |
| Low mental labor job | 129 (35.1) | 135 (37.2) | 264 (36.1) | 320 (33.5) |
| High mental labor job | 88 (23.9) | 86 (23.7) | 174 (23.8) | 295 (30.9) |
| Left ventricular wall score | 7.44 (2.89) | 7.30 (2.78) | 7.37 (2.84) | 7.07 (2.69) |
| Taken nitrates regularly | 205 (55.7) | 196 (54.0) | 401 (54.9) | 528 (55.3) |
| History of MI | 209 (56.8) | 228 (62.8) | 437 (59.8) | 549 (57.5) |
| Male | 35 (9.5) | 37 (10.2) | 72 (9.8) | 87 (9.1) |
| Smoking status | | | | |
| Never smoked | 62 (16.8) | 54 (14.9) | 116 (15.9) | 157 (16.4) |
| Former smoker | 164 (44.6) | 157 (43.3) | 321 (43.9) | 451 (47.2) |
| Current smoker | 142 (38.6) | 152 (41.9) | 294 (40.2) | 347 (36.3) |
| High limitation of activities | 165 (44.8) | 173 (47.7) | 338 (46.2) | 441 (46.2) |
| High recreational activity | 228 (62.0) | 219 (60.3) | 447 (61.1) | 616 (64.5) |
| Confirmed hypertension | 118 (32.1) | 108 (29.8) | 226 (30.9) | 260 (27.2) |
| Diabetes status | | | | |
| No diabetes | 325 (88.3) | 328 (90.4) | 653 (89.3) | 873 (91.4) |
| Uncertain diabetes | 13 (3.5) | 7 (1.9) | 20 (2.7) | 23 (2.4) |
| Confirmed diabetes | 30 (8.2) | 28 (7.7) | 58 (7.9) | 59 (6.2) |
| LMCA percent obstruction | 4.27 (11.87) | 2.78 (9.55) | 3.53 (10.80) | 5.76 (14.50) |
| PLMA percent obstruction | 36.44 (38.04) | 34.89 (36.95) | 35.67 (37.49) | 39.14 (38.73) |
| Any diseased proximal vessels | 222 (60.3) | 230 (63.4) | 452 (61.8) | 608 (63.7) |
| Systolic blood pressure | 130.28 (17.40) | 130.34 (18.72) | 130.31 (18.06) | 129.80 (18.23) |

LMCA = left main coronary artery; MI = myocardial infarction; PLMA = proximal left anterior artery
For continuous variables we report the mean (standard deviation); for binary variables we report the number of patients (percentage)

variable) were indeed effect modifiers in the target population of all trial-eligible patients.

A total of 1686 patients had complete data on the baseline covariates we used in our analysis (731 randomized, 368 to surgery and 363 to medical therapy; 955 non-randomized, 430 receiving surgery and 525 medical therapy). Table 1 summarizes the basic descriptive statistics for the baseline covariates. In general, randomized and non-randomized patients had similar characteristics, but non-randomized patients were more likely to have taken a beta-blocker regularly, have a higher left main coronary percent obstruction, and have a higher left ventricular wall score.

## Statistical methods

We evaluated the target population CATEs (risk differences) for mortality at 10 years of follow-up, conditional on history of myocardial infarction and baseline ejection fraction value. We analyzed the 986 patients with a history of myocardial infarction and the 700 patients without a history of myocardial infarction separately, when estimating the nuisance functions and when estimating the pseudo-outcome regression. We obtained pointwise confidence intervals and uniform confidence bands within the subgroups defined by history of myocardial infarction.

In the first step to estimate the nuisance functions for the outcome and participation, we used parametric models (logistic regression) and included the main effects of all baseline covariates listed in Table 1, except we modeled age

and ejection fraction using B-splines (basis splines) of order 3 (degree 2) with an interior knot placed at the median of age or ejection fraction. We modeled the outcome separately in each treatment group in the trial. Because the model for treatment in the trial cannot be misspecified, to estimate $e_a(X)$, we used a simple logistic model that included the main effects of age, severity of angina, ejection fraction value, systolic blood pressure, proximal left anterior artery percent obstruction, and left ventricular wall score [6].

In the second step, we fit the regression of the pseudo-outcome on ejection fraction. We used ordinary least squares regression with either a B-spline or polynomial of ejection fraction. We did not want to assume that the CATE function has a linear form, so we chose to use B-splines of order 3 with an interior knot placed at the median of ejection fraction. For the polynomial of ejection fraction, we set the degree to 3. These models are flexible enough to capture most reasonable nonlinearities in the treatment effect over ejection fraction. When forming the pointwise confidence intervals, we used the robust variance estimator. We obtained uniform confidence bands using the weighted bootstrap procedure with 200 replicates.

To evaluate the robustness of our results to model specification of the nuisance functions, we repeated our analyses using generalized additive models (GAMs) instead of parametric models in the first step of the procedure. For comparison, we also estimated trial-only CATEs, by modifying the procedure to only use the trial pseudo-outcome (which does not include the participation weight) and to fit the second step regression only among trial-participants [18].

## Results

Figure 1 shows the estimated target population CATE function stratified by history of myocardial infarction over a set of ejection fraction values, from 40 to 80, when using parametric models in the first step and splines in the second step. The CATE functions for patients with and without a history of myocardial infarction show different patterns, but both look like they could be reasonably well-approximated by a linear fit. For patients with a history of myocardial infarction, the CATE function linearly increases from a risk difference of approximately -0.25 for patients with an ejection fraction of 40% up to a risk difference of approximately 0.15 for patients with an ejection fraction of 80%. The uniform confidence band suggests that the data are incompatible with the hypothesis that the CATE function is constant at 0 (no effect), over the set of ejection fraction values we considered. In other words, among patients with a history of myocardial infarction, the treatment effect appears to vary over ejection fraction, suggesting that surgery may be more beneficial (compared to medical therapy) for patients with lower ejection fraction, compared to those with higher ejection fraction.

For patients without a history of myocardial infarction, the CATE function decreases from a risk difference of approximately 0.15 for patients with an ejection fraction of 40% to slightly less than 0 for patients with an ejection fraction of 70%; then it increases up to approximately 0.20 for patients with an ejection fraction of 80%. Because the uniform confidence band contains zero, across all levels of ejection fraction, the data are not incompatible with the

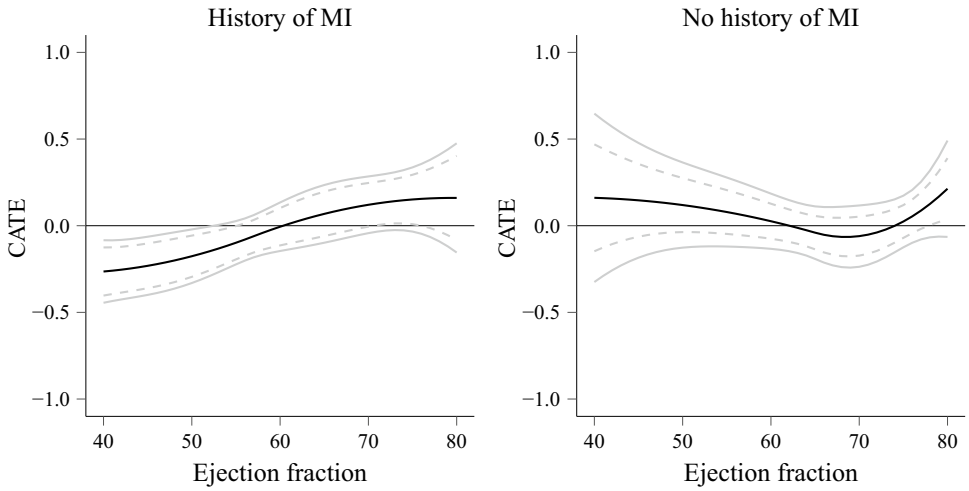


**Fig. 1** Target population CATE function estimated using parametric models in the first step and spline regression in the second step CATE = conditional average treatment effect; MI = myocardial infarction. The black line indicates the estimated CATE function; dashed gray lines connect 95% pointwise upper and lower confidence limits; solid gray lines depict the uniform 95% confidence band. The set of grid points went up to ejection fraction values of 80%, over a grid of evenly spaced points in steps of 1%. In each panel, the confidence bands are uniform over ejection fraction (conditional on history of MI).

hypothesis that the CATE function is constant over ejection fraction for patients without a history of myocardial infarction.

In Web Appendix 6, we provide additional results for the CASS analysis. Web Appendix Figure 1 shows that the second step regression using polynomials instead of splines yielded similar results. The estimated trial-only CATE functions, provided in Web Appendix Figures 2 and 3, were similar to the corresponding estimated target population CATE functions. We also found that repeating the analysis with GAMs in the first step resulted in similar CATE functions; see Web Appendix Figures 4 through 7. We have provided a simulated dataset and R code [63] that implements the two-step estimation procedure and produces graphs of the CATE function on GitHub (https://github.com/serobertson/GeneralizabilityCATE). We provide additional details about the code in Web Appendix 7.

## Discussion

We described a two-step pseudo-outcome regression procedure for estimating target population CATEs in nested trial designs used to extend inferences from a randomized trial to a target population. We also described how to obtain pointwise confidence intervals for the CATE at specific effect modifier values and uniform confidence bands for the CATE function. This two-step procedure provides a regression-based framework for examining CATEs given discrete as well as continuous covariates, whereas previously proposed methods only allow the estimation of CATEs within subgroups defined by discrete covariates [12, 13]. Even when all covariates of interest are discrete, working within a regression framework may be advantageous because it allows the representation of smoothness or homogeneity assumptions by omitting covariate-by-covariate product terms from the regression specification; such assumptions are not as easy to represent with previously proposed methods [12, 13].

We note the different roles of the baseline covariates in the two steps of the procedure: the first step "controls" for enough variables to address selective trial participation; the second step focuses on a much smaller set of key effect modifiers. This duality is analogous to the difference between the variables needed to address confounding and effect modifiers in previous work on estimating CATEs in observational studies [15–21, 64]. In applications, examining heterogeneity over a lower dimensional set of covariates may be motivated by scientific or policy considerations. For example, key effect modifiers may be identified on the basis of previous investigations, and the methods described here can be used in confirmatory assessments of heterogeneity. Or, it may be desirable to base treatment decisions on only a subset of potential effect modifiers while ignoring unacceptable ones (e.g., even if insurance status were a strong effect modifier, we might prefer to not use it to make treatment decisions; instead, we might choose to examine heterogeneity only over lab measurements or past medical history).

Our methods are motivated by applications in which a few key effect modifiers of interest can be identified by the investigators (e.g., on the basis of prior studies). When the effect modifiers of interest are more numerous it may be possible to summarize them into a score (e.g., using an outcome or treatment effect model obtained from external data) and use that score in the second step of our procedure. It is also possible to extend our methods to settings where $\widetilde{X}$ is of moderate to high dimensionality, or even substituting $X$ for $\widetilde{X}$, as is often the case in discovery-oriented investigations [18, 65–67] (we briefly touch on these approaches in Web Appendix 1 and Web Appendix 5). In such investigations, the study goal is to predict individualized responses for members of the target population and the second step of the procedure typically is modified to use data-adaptive approaches appropriate for high-dimensional covariates [18]. The development of methods for valid inference in this context is an area of active research. Broadly speaking, when examining heterogeneity over high-dimensional covariates, there exists a trade-off between the flexibility of the model specification and the strength of the technical assumptions needed for valid estimation and inference [68].

In summary, we proposed a two-step estimation procedure for estimating the target population CATE as a function of key effect modifiers in nested trial designs. This procedure is useful for examining the dependence of the CATE on a small set of key effect modifiers, while adjusting for a large set of covariates needed to ensure the exchangeability of the trial and the target population.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10654-022-00901-5.

**Author contributions** All authors were involved in drafting the manuscript and have read and approved the final version submitted. SER conducted the statistical analysis.

**Funding** This work was supported in part by Agency for Healthcare Research and Quality (AHRQ) award R36HS028373-01 and Patient-Centered Outcomes Research Institute (PCORI) awards ME-1502-27794, ME-2019C3-17875, and ME-2021C2-22365, and National Library of Medicine (NLM) award R01LM013616. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of PCORI, the PCORI Board of Governors, or the PCORI Methodology Committee, NLM, or the CASS investigators.

**Data Availability** The analyses in our paper used CASS research materials obtained from the National Heart, Lung, and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center.

**Code availability** Code is available online at GitHub [https://github.com/serobertson/GeneralizabilityCATE].

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Ethical approval** Our research is not human subjects research.

## References


1. Hernán MA. "Discussion of "Perils and potentials of self-selected entry to epidemiological studies and surveys. J Royal Stat Soc Series A (Statistics in Society). 2016;179(2):346–7.
2. Dahabreh IJ, Hernán MA. Extending inferences from a randomized trial to a target population. Eur J Epidemiol. 2019;34(8):719–22.
3. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. Am J Epidemiol. 2010;172(1):107–15.
4. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2017;186(8):1010–4.
5. Rudolph KE, van der Laan MJ. Robust estimation of encouragement design intervention effects transported across sites. J Royal Stat Soc Series B (Statistical Methodology). 2017;79(5):1509–25.
6. Dahabreh IJ, Robertson SE, Tchetgen Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics. 2018;75(2):685–94.
7. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020;39(14):1999–2014.
8. 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.
9. Seamans MJ, Hong H, Ackerman B, Schmid I, Stuart EA. Generalizability of subgroup effects. Epidemiology. 2021;32(3):389–92.
10. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology. 2007;18(5):561–8.
11. Hernán MA, Robins JM. Causal Inference: What If. 1st ed. Boca Raton, FL: Chapman & Hall/CRC; 2020.
12. Mehrotra ML, Westreich D, Glymour MM, Geng E, Glidden DV. Transporting subgroup analyses of randomized trials for planning implementation of new interventions'. Am J Epidemiol. 2021;190(8):1671–80.
13. Robertson SE, Steingrimsson JA, Joyce NR, Stuart EA, Dahabreh IJ. Estimating subgroup effects in generalizability and transportability analyses," American Journal of Epidemiology, kwac036, 2022.
14. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Stat Med. 1997;16(3):285–319.
15. Abrevaya J, Hsu Y-C, Lieli RP. Estimating conditional average treatment effects. J Bus Econom Stat. 2015;33(4):485–505.
16. Lee S, Okui R, Whang Y-J. Doubly robust uniform confidence band for the conditional average treatment effect function. J Appl Econom. 2017;32(7):1207–25.
17. Lechner M. Modified causal forests for estimating heterogeneous causal effects. arXiv preprint arXiv:1812.09487, 2018.
18. Kennedy EH. Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint arXiv:2004.14497, 2020.
19. Semenova V, Chernozhukov V. Debiased machine learning of conditional average treatment effects and other causal functions. Econom J. 2021;24(2):264–89.
20. Fan Q, Hsu Y-C, Lieli RP, Zhang Y. Estimation of conditional average treatment effects with high-dimensional data. J Bus Econom Stat. 2020;40(1):313–27.
21. Knaus MC, Lechner M, Strittmatter A. Machine learning estimation of heterogeneous causal effects: empirical monte carlo evidence. Econom J. 2021;24(1):134–61.
22. Dahabreh IJ, Haneuse SJ-P, Robins JM, Robertson SE, Buchanan AL, Stuart EA, Hernán MA. Study designs for extending causal inferences from a randomized trial to a target population. Am J Epidemiol. 2021;190(8):1632–42.
23. Robins JM. Confidence intervals for causal parameters. Stat Med. 1988;7(7):773–85.
24. Splawa-Neyman J. On the application of probability theory to agricultural experiments. essay on principles. section 9. [Translated from Splawa-Neyman, J (1923) in Roczniki Nauk Rolniczych Tom X, 1–51]. Stat Sci. 1990;5(4):465–72.
25. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
26. Robins JM, Greenland S. Causal inference without counterfactuals: comment. J Am Stat Assoc. 2000;95(450):431–5.
27. CASS Principal Investigators. Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery: comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria. J Am Collegef Cardiol. 1984;3(1):114–28.
28. Passamani E, Davis KB, Gillespie MJ, Killip T, Investigators CP, Associates T. A randomized trial of coronary artery bypass surgery: survival of patients with a low ejection fraction. New England J Med. 1985;312(26):1665–71.
29. Dahabreh IJ, Robins JM, Haneuse SJ-P, Hernán MA. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. arXiv preprint arXiv:1906.10792, 2019 (accessed: 11/03/2020).
30. Rubin DB. Statistics and causal inference: Comment: Which ifs have causal answers. J Am Stat Assoc. 1986;81(396):961–2.
31. Rubin DB. Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes. Psychol Method. 2010;15(1):38–46.
32. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009;20(6):880–3.
33. Halloran ME, Struchiner CJ. Causal inference in infectious diseases. Epidemiology, 1995; pp. 142–151. https://pubmed.ncbi.nlm.nih.gov/7742400.
34. Dahabreh IJ, Robins JM, Haneuse SJ-P, Saeed I, Robertson SE, Stuart EA, Hernán MA. "Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population," arXiv preprint arXiv:1905.10684, 2019.
35. Pearl J, Bareinboim E. Transportability of causal and statistical relations: A formal approach. In: 11th AAAI conference on artificial intelligence 2011 Aug 4 pp. 540–547.
36. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Method Med Res. 2012;21(1):31–54.
37. Robins JM, Hernán MA. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis* G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, eds.) (pp. 567-614). Chapman and Hall/CRC.
38. Dahabreh IJ, Robins JM, Hernán MA. Benchmarking observational methods by comparing randomized trials and their emulations. Epidemiology. 2020;31(5):614–9.
39. Tsiatis A. *Semiparametric theory and missing data*. New York:Springer, 2007. https://link.springer.com/book/10.1007/0-387-37345-4.

40. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1–68.
41. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–73.
42. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–60.
43. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat Med. 2014;33(5):721–37.
44. Racine JS. *Nonparametric Econometrics: A Primer*. Foundation and Trends in Econometrics, 2008.https://socialsciences.mcmaster.ca/racinej/ECO0301.pdf.
45. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton, FL: Chapman & Hall/CRC; 2020.
46. Smucler E, Rotnitzky A, Robins JM. "A unifying approach for doubly-robust $\ell_1$ regularized estimation of causal contrasts," arXiv preprint arXiv:1904.03737, 2019.
47. Benkeser D, Van Der Laan M. "The highly adaptive lasso estimator," In :2016 IEEE international conference on data science and advanced analytics (DSAA), pp. 689–696, IEEE, 2016.
48. Horowitz JL. *Semiparametric and nonparametric methods in econometrics*. New York: Springer, 2009. https://link.springer.com/book/10.1007/978-0-387-92870-8.
49. Kennedy EH, Lorch S, Small DS. Robust causal inference with continuous instruments using the local instrumental variable curve. J Royal Stat Soc: Series B (Statistical Methodology). 2019;81:121–43.
50. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
51. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc: Series B (Statistical Methodology). 1996;58(1):267–88.
52. Efron B, Tibshirani RJ. *An introduction to the bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 1994.
53. Huber PJ. Under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; University of California Press: Berkeley, CA, USA 1967 (p. 221).
54. Stefanski LA, Boos DD. The calculus of M-estimation. Am Stat. 2002;56(1):29–38.
55. Belloni A, Chernozhukov V, Chetverikov D, Kato K. Some new asymptotic theory for least squares series: pointwise and uniform results. J Econom. 2015;186(2):345–66.
56. Belloni A, Chernozhukov V, Chetverikov D, Wei Y. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. Annal stat. 2018;46(6B):3643.
57. Vaart AW, Wellner JA. Weak convergence. InWeak convergence and empirical processes 1996 (pp. 16-28). Springer, New York, NY.
58. William J, Russell R, Nicholas T, et al. Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Circulation. 1983;68(5):939–50.
59. Alderman EL, Bourassa MG, Cohen LS, Davis KB, Kaiser GG, Killip T, Mock MB, Pettinger M, Robertson T. Ten-year follow-up of survival and myocardial infarction in the randomized coronary artery surgery study. Circulation. 1990;82(5):1629–46.
60. Robertson SE, Leith A, Schmid CH, Dahabreh IJ. Assessing heterogeneity of treatment effects in observational studies. Am J Epidemiol. 2021;190(6):1088–100.
61. Yusuf S, Zucker D, Passamani E, Peduzzi P, Takaro T, Fisher L, Kennedy J, Davis K, Killip T, Norris R, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the coronary artery bypass graft surgery trialists collaboration. The Lancet. 1994;344(8922):563–70.
62. Velazquez EJ, Lee KL, Jones RH, Al-Khalidi HR, Hill JA, Panza JA, Michler RE, Bonow RO, Doenst T, Petrie MC, et al. Coronary-artery bypass surgery in patients with ischemic cardiomyopathy. New England J Med. 2016;374(16):1511–20.
63. Core Team R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021.
64. Zimmert M, Lechner M. "Nonparametric estimation of causal heterogeneity under high-dimensional confounding," arXiv preprint arXiv:1908.08779, 2019.
65. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Nat Acad Sci. 2019;116(10):4156–65.
66. Nie X, Wager S. "Quasi-oracle estimation of heterogeneous treatment effects," arXiv preprint arXiv:1712.04912, 2017.
67. Athey S, Wager S. "Estimating treatment effects with causal forests: An application," arXiv preprint arXiv:1902.07409, 2019.
68. Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I. "Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in India," National Bureau of Economic Research, 2018. https://arxiv.org/abs/1712.04802.