# Generalizability Methods for Estimating Causal Population Effects

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

# DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

**Department of Biostatistics**

have examined a dissertation entitled

"Generalizability Methods for Estimating Causal Population Effects"

presented by Irina Degtiar

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

*Signature* Francesca dominici .....

*Typed name*: Prof. Francesca Dominici

*Signature* Sherri Rose
Sherri Rose (May 6, 2021 18:38 PDT) .....

*Typed name*: Prof. Sherri Rose

*Signature* .....

*Typed name*: Prof. Sebastien Haneuse

*Signature* .....

*Typed name*:

*Date*: May 6, 2021

# Generalizability Methods for Estimating Causal Population Effects

A dissertation presented
by

Irina Degtiar

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2021

Dissertation Advisors: Professor Sherri Rose,
Professor Francesca Dominici

Irina Degtiar

# Generalizability Methods for Estimating Causal Population Effects

## Abstract

Studies are often performed in samples that do not resemble the target populations relevant for policy, treatment, or other decisions. Much of the causal inference literature has focused on addressing internal validity bias; however, both internal and external validity are necessary for unbiased estimates in a target population. The generalizability methods presented in this thesis allow for inference on the population of interest rather than the one in the study.

Chapter 1 presents a framework for addressing external validity bias, including a synthesis of approaches for generalizability and transportability, the assumptions they require, as well as tests for the heterogeneity of treatment effects and differences between study and target populations. The chapter concludes with practical guidance for researchers and practitioners.

Chapter 2 presents an innovative class of estimators, conditional cross-design synthesis (CCDS), for combining randomized and observational data to eliminate their respective external and internal validity biases. CCDS uses the region of covariate overlap between data types to remove potential unmeasured confounding bias in the observational data in order to extend inference beyond the support of the randomized data to the full target population. We derive outcome regression, propensity weighting, and double robust approaches under the CCDS framework. We illustrate the methods to estimate the causal effect of health insurance plans on cost among New York City Medicaid enrollees.

Chapter 3 introduces novel approaches for generalizing from an evaluation study of a voluntary intervention to estimate population average treatment effects for future treated individuals, which can accommodate nonparametric outcome regression approaches such

as Bayesian Additive Regression Trees and Bayesian Causal Forests. The generalizability approach incorporates uncertainty regarding target population treated group membership into the posterior credible intervals to better-reflect the uncertainty of scaling up a voluntary intervention. In a simulation based on real data, we estimate impacts of a national scale-up of a voluntary health policy model and highlight the benefit of using flexible regression approaches for generalizability.

# Contents

# List of Tables

# List of Figures

# Acknowledgments

This thesis would not have been possible without the support and encouragement from the community around me. I would like to extend my infinite gratitude to my advisor Sherri Rose, who believed in me and encouraged me, particularly when I was doubting my work the most. Thank you for supporting me 100% not just in my research but also in my career, interests, and beyond throughout these years, and for helping me recognize and celebrate my accomplishments, both professional and otherwise.

I was also incredibly lucky to be able to collaborate with my co-adviser Francesca Dominici and my committee member Sebastien Haneuse. Many thanks to Francesca, who helped me to step back and remember the bigger picture of where my projects were headed when I bogged myself down in the details. Thank you to Sebastien – whose on-point questions I have admired since taking his Analysis of Multivariate and Longitudinal Data class – for always forcing me to dig deeper and keep questioning the "why."

Chapter 3 of my dissertation would not have been possible without an opportune seminar Mariel Finucane gave during my internship at Mathematica Policy Research, which launched off a collaboration that led to this work. Thank you, Mariel, for being the most supportive, sweet, and genuinely positive person I have had the pleasure to work with.

I would not have gotten through this program without my brilliant classmates and cohort-mates: thank you for your insightful questions, problem set support, collegiality, and willingness to go on hiking adventures together. I am likewise grateful to the Health Policy Data Science lab for providing community and feedback; you have only grown stronger in now being a dual-institution lab! Thank you also to the fountain of infinite department wisdom, Jelena Follweiler, and to other department faculty and staff for helping me throughout my journey over the last 5 years and shaping me into the statistician I am today.

My sanity throughout these years would not have been maintained without the generous support of my husband, Anshuman, who has been a solid bedrock through my waves of stress over these years. Thank you for selflessly putting aside your own work to help support me in every way you can, from cooking to coding. I am also so grateful to my Boston dance

communities and hiking buddies, who have provided me a joyous outlet and reminded me that life is more than just work. Particularly during this last pandemic year, I am indebted to the virtual hangouts, workouts, and calls with my friends from around the country and the world, which have reminded me to come up for a breath of air every now and then in the midst of my thesis work.

Last but not least, I would not be here today without my family, who have provided me with unwavering love and support. Thank you for always being by my side, even when you are on the other side of the country.

# Chapter 1

# A Review of Generalizability and Transportability

Irina Degtiar[1], Sherri Rose [2]

[1] *Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA*

[2] *Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, CA, USA*

## Abstract

When assessing causal effects, determining the target population to which the results are intended to generalize is a critical decision. Randomized and observational studies each have strengths and limitations for estimating causal effects in a target population. Estimates from randomized data may have internal validity but are often not representative of the target population. Observational data may better reflect the target population, and hence be more likely to have external validity, but are subject to potential bias due to unmeasured confounding. While much of the causal inference literature has focused on addressing internal validity bias, both internal and external validity are necessary for unbiased estimates in a target population. This paper presents a framework for addressing external validity bias, including a synthesis of approaches for generalizability and transportability, the assumptions they require, as well as tests for the heterogeneity of treatment effects and differences between study and target populations.

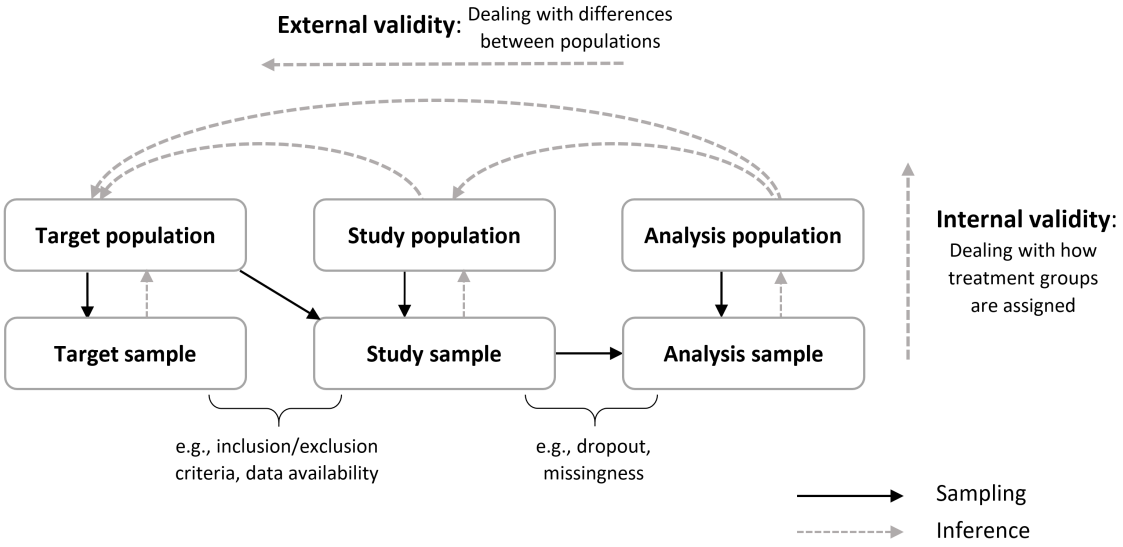



**Figure (1.1)** *Internal vs. external validity biases as they relate to target, study, and analysis populations*

## 1.1 Background

The goal of causal inference is often to gain understanding of a particular target population based on study findings. The true underlying causal effect will typically vary with the definition of the chosen target population. However, samples unrepresentative of the target population arise frequently in studies ranging from randomized controlled trials (RCTs) in clinical medicine to policy research (Bell *et al.*, 2016; Kennedy-Martin *et al.*, 2015; Allcott, 2015). In a clinical trial setting, physicians may be left interpreting evidence from RCTs with patients who have demographics and comorbidities that are quite different from those of their patients. As an example, within cancer RCTs, African Americans are widely underrepresented despite being at an increased risk for many cancers (Chen and Wong, 2018). Failing to address this lack of representation can lead to inappropriate conclusions and harm (Chen *et al.*, 2020). In a policy setting, it is important to consider the effects that can be expected in the eventual target population in order to set expectations for anticipated results and determine groups that should be targeted for an intervention.

The relationships between target, study, and analysis populations are visualized in Figure 1.1. The target sample is a representative sample of the target population, whereas

the study population is defined by enrollment processes and inclusion or exclusion criteria. Due to these practical and scientific considerations, the study population may differ from the target population. Correspondingly, the enrolled participants who form the study sample may have different characteristics from those of the target sample. In the cancer RCT example, while a physician might care about the target population of patients that may come in to be treated by their clinic (of which the clinic's current patients are a target sample), the study sample on which they're basing their treatment recommendations may not include any African Americans. The study population is the hypothetical population that the study sample represents, which likewise includes no African Americans. Post-enrollment, further dropout and missingness may occur that create the observed analysis sample. In this case, dropout may have occurred for patients who experienced severe adverse events such that the analysis sample consists of patients who did not experience severe side effects. There then exists a hypothetical analysis population from which the analysis sample data is a simple random sample. Hereafter, for simplicity and consistency with the literature, we will use the terms study sample and study population to be inclusive of the analysis sample and analysis populations, respectively.

Several key concepts are crucial to understand when considering extending causal inferences beyond a study sample. *Generalizability* focuses on the setting where the study population is a subset of the target population of interest, while *transportability* addresses the setting where the study population is (at least partly) external to the target population. *Internal validity* is defined as an effect estimate being unbiased for the causal treatment effect in the population from which the sample is a simple random sample (i.e., moving vertically from a sample to its corresponding population in Figure 1.1). *External validity* is concerned with how well results generalize to other contexts. Specifically, that the (internally valid) effect estimate is unbiased for the causal treatment effect in a different setting, such as a target population of interest (moving laterally between populations in Figure 1.1). External validity bias has also been referred to as sample selection bias (Heckman, 1979; Imai *et al.*, 2008; Moreno-Torres *et al.*, 2012; Bareinboim *et al.*, 2014; Haneuse, 2016).

External validity bias arises from differences between the study and target populations in (1) subject characteristics; (2) setting, such as geography or type of health center; (3) treatment, such as timing, dosage, or staff training; and (4) outcomes, such as length of follow-up or timing of measurements (Cronbach and Shapiro, 1982; Rothwell, 2005; Dekkers *et al.*, 2010; Green and Glasgow, 2006; Burchett *et al.*, 2011; Attanasio *et al.*, 2003). The focus of most generalizability and transportability methods is on addressing differences in subject characteristics. Hence, these methods assume the remaining threats to external validity are not present in the data sources they are looking to generalize across. Namely, external validity bias then arises solely from: (1) variation in the probability of enrollment in the study, (2) heterogeneity in treatment effects, and (3) the correlation between (1) and (2) (Olsen *et al.*, 2013). We therefore distinguish between factors differentiating the target population from the study population (external validity bias) and those that create differences between treatment groups (internal validity bias), e.g., confounding. RCTs are frequently performed in a nonrepresentative subset of the target population and may have imperfect follow-up (challenging their external validity) and may have baseline imbalances (leading to internal validity bias). Observational studies may be susceptible to unmeasured confounding (threatening their internal validity), but may be more representative of the target population (hence having better external validity). Lack of representation in an RCT can lead to external validity bias that is larger than the internal validity bias of an observational study (Bell *et al.*, 2016).

The optimal solution to external validity bias centers on study design, which we review briefly here, but do not cover extensively. One type of ideal study would randomly sample subjects from the target population and then randomly assign treatment to the selected individuals. However, this is usually infeasible. Alternative study designs for improving study generalizability and transportability include purposive sampling, where investigators deliberately select individuals such as for representation or heterogeneity (Shadish *et al.*, 2001; Allcott and Mullainathan, 2012); pragmatic or practical clinical trials, which aim to be representative of clinical practice (Schwartz and Lellouch, 1967; Ford and Norrie, 2016);

stratified selection based on effect modifiers or propensity scores for selection (Tipton *et al.*, 2014; Tipton, 2013b; Allcott and Mullainathan, 2012); and balanced sampling designs for site selection that select representative sites through stratified ranked sampling (Tipton and Peck, 2017). In lieu of or in addition to study designs that address external validity bias, generalizability and transportability methods can improve the external validity of effect estimates after data collection.

This manuscript provides a review of generalizability and transportability research, synthesizing across the statistics, epidemiology, computer science, and economics literature in a more complete manner than has been done to date. Existing review literature has examined narrower subsets of the topic: generalizing or transporting to a target population from only RCT data (Stuart *et al.*, 2015, 2018; Kern *et al.*, 2016; Tipton and Olsen, 2018; Ackerman *et al.*, 2019), identifiability rather than estimation (Bareinboim and Pearl, 2016), or meta-analysis approaches for combining summary-level information (Verde and Ohmann, 2015; Kaizar, 2015). A recent related review on combining randomized and observational data featured a simulation, real data analysis, and software guide (Colnet *et al.*, 2020). However, these previous reviews have not summarized the full range of generalizability and transportability methods that incorporate data from randomized, observational, or a combination of randomized and observational studies, nor techniques for evaluating generalizability, as we do here. Additionally, although the importance of describing generalizability and transportability is recognized by different trial reporting guidelines (e.g., CONSORT, RECORD, STROBE), they provide no clear guidance on tests or estimation procedures (Schulz *et al.*, 2010; Benchimol *et al.*, 2015; von Elm *et al.*, 2008). We also contribute recommendations for methodologists and applied researchers.

The remainder of the article synthesizes considerations for assessing and addressing external validity bias after data collection (presented as a framework in Figure 2) and is organized as follows. Section 2 defines the estimand of interest, the average treatment effect in a target population, as well as alternatives. Section 3 presents key assumptions underlying many of the methods. Section 4 reviews methods for assessing treatment effect heterogeneity,

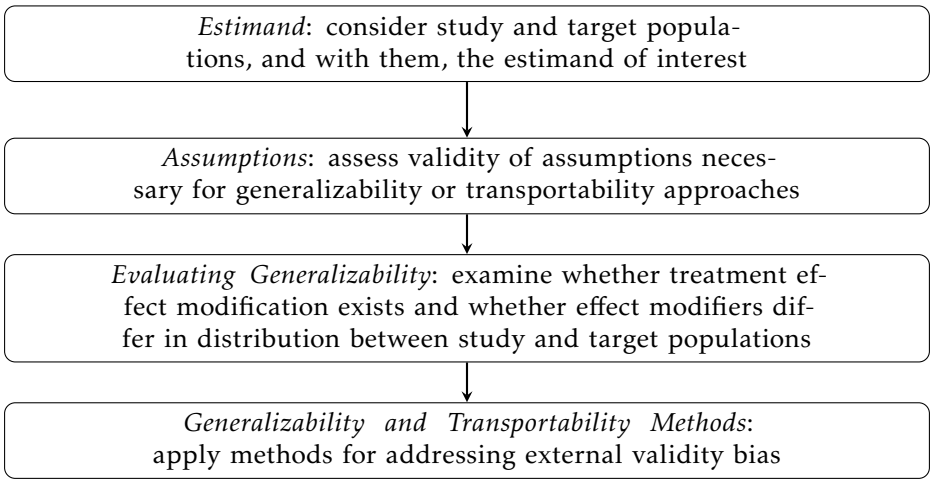



**Figure (1.2)** *Overview framework for assessing and addressing external validity bias after data collection*

thus further motivating the need for methods that enable generalizing or transporting study results to a target population. Section 5 then summarizes the analytic methods available for external validity bias correction that generate treatment effect estimates for a target population of interest. These techniques include weighting and matching, outcome regressions, and doubly robust approaches. Section 6 then concludes with guidance for both applied and methods researchers.

## 1.2 Estimand

Assume, for one or more studies, the existence of outcome $Y$, treatment $A \in \{0,1\}$, and baseline covariates $X \in \mathbb{R}^d$. For simplicity of notation, we define $X$ to represent all treatment effect confounders and effect modifiers (subgroups whose effects are expected to differ) that differ between study and target populations; each variable in $X$ is both a confounder and an effect modifier. Without loss of generality, we focus on the single study setting, with $S = 1$ indicating selection into it. The observational unit for the study sample is $O_{\text{study}} = \{X, A, Y, S = 1\}$. $O_{\text{study}}$ has probability distribution $P_{\text{study}} \in \mathcal{M}_{\text{study}}$, where $\mathcal{M}_{\text{study}}$ is our collection of possible probability distributions (i.e., statistical model). We observe $n_s$ realizations of $O_{\text{study}}$, indexed by $j$. The observational unit for a representative sample from the target population is given by $O = \{X, A, Y, S\} \sim P \in \mathcal{M}$. We observe $n$ realizations of

$O$, indexed by $i$. Target sample subjects who do not appear in the study sample will have $S = 0$. We use the terminology "selected" or "sampled" throughout the paper for simplicity although for transportability, subjects are not directly sampled into the study from the target population. For generalizability, $O_{\text{study}} \in O$, while for transportability, the two are disjoint sets, $O_{\text{study}} \notin O$.

Biases are defined with respect to an estimand. We will focus on the average treatment effect in a well-defined target population of interest: the population average treatment effect (PATE). Namely, we are interested in the average outcome had everyone in the target population been assigned to treatment *A=1* compared to the outcome had everyone been assigned to treatment *A=0*. We write this as $\tau = E_X(E(Y|S = 1, A = 1, X) - E(Y|S = 1, A = 0, X)) = E(Y^1 - Y^0)$, where $Y^1$ and $Y^0$ are the potential outcomes under treatment and no treatment, respectively, and required identifiability assumptions are delineated in the next section. The corresponding estimator is given by $\hat{\tau} = 1/n \sum_{i=1}^{n} (\hat{Y}_i^1 - \hat{Y}_i^0)$. We also write $Y^a$ to represent the potential outcome under $a$ with lowercase $a$ a specific value for random variable $A$. Potential outcomes are either explicitly assumed in the potential outcomes framework or a consequence of the structural causal model (Rubin, 1974; Pearl, 2000). Different target populations correspond to alternative PATEs because the expectation is taken with respect to alternative distributions of covariates $X$. However, necessarily, we only observe outcomes in the study sample. A study therefore directly estimates the sample average treatment effect (SATE): $\tau_s = E(Y^1 - Y^0|S = 1)$ with estimator $\hat{\tau}_s = 1/n_s \sum_{j:S_j=1} (\hat{Y}_j^1 - \hat{Y}_j^0)$.

When the distributions of treatment effect modifiers differ between study and target populations, the true study average effect will not equal the true target population average effect (SATE $\neq$ PATE) due to external validity bias. Sampling variability as well as internal validity biases can also drive estimates of SATE further from the truth (Figure 1.3). Biases may differ in magnitude and may make the SATE either larger or smaller than the PATE.

We may also be interested in estimating other target parameters. For example, the population conditional average treatment effects (PCATE): $\tau_x = E(Y^1 - Y^0|X)$ is examined in some of the estimation methods we explore later. Another parameter of interest is the

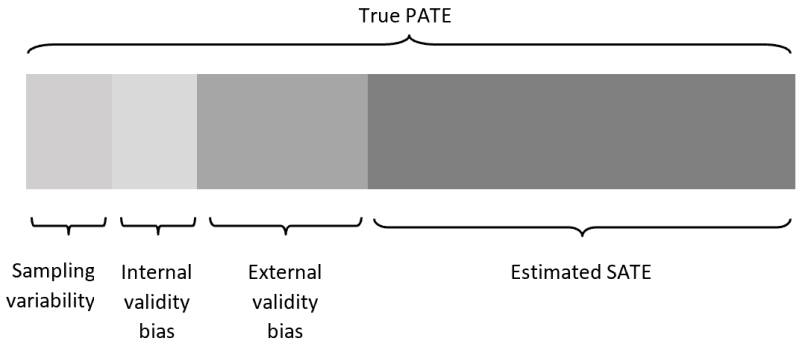



**Figure (1.3)** *Illustrative example of the difference between target population and sample average treatment effects (PATE and SATE)*

*Biases may differ in magnitude and may make the SATE either larger or smaller than the PATE.*

population average treatment effects among the treated: $\tau_1 = E(Y^1 - Y^0|A = 1)$. Similar generalizability and transportability considerations presented in the following sections will apply for these and other causal estimands.

## 1.3 Assumptions

Under the potential outcomes framework, the assumptions below are sufficient to identify the PATE using the observed study data. A corresponding set of assumptions under the structural equation model (SEM) framework has also been derived (Pearl and Bareinboim, 2014; Pearl, 2015; Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2014; Bareinboim and Tian, 2015; Bareinboim and Pearl, 2016; Correa *et al.*, 2018). Additional assumptions include those of no missing data or measurement error in outcome, treatment, or covariate measurements. Other target parameters of interest necessitate a similar set of assumptions.

### 1.3.1 Internal validity

Sufficient assumptions for identifying the PATE with respect to internal validity:

**Conditional treatment exchangeability**: $Y^a \perp A \,|X, S = 1$ for all $a \in \mathcal{A}$, the set of all possible treatments. This condition requires no unmeasured confounding of the treatment-outcome relationship in the study. It is satisfied by perfectly randomized trials (e.g., no loss to

follow-up, other informative missingness or censoring, etc.) and by observational studies that have all confounders measured. While this condition is sufficient, it is not always necessary. When estimating the PATE, it can be replaced by the weaker condition of mean conditional exchangeability of the treatment effect, $E(Y^1 - Y^0|X, A, S = 1) = E(Y^1 - Y^0|X, S = 1)$ (Kern *et al.*, 2016; Dahabreh *et al.*, 2019c).

**Positivity of treatment assignment**: $P(X = x|S = 1) > 0 \Rightarrow P(A = a|X = x, S = 1) > 0$, with probability 1 for all $a \in \mathcal{A}$. This condition entails that each subject in the study has a positive probability of receiving each version of the treatment. In combination with the conditional treatment exchangeability assumption above, this assumption is also known as strongly ignorable treatment assignment (Varadhan *et al.*, 2016).

**Stable unit treatment value assumption (SUTVA) for treatment assignment**: if $A = a$ then $Y = Y^a$. This assumption requires no interference between subjects and treatment version irrelevance (i.e., consistency/well-defined interventions) in the study and target populations (Dahabreh *et al.*, 2017; Kallus *et al.*, 2018).

### 1.3.2 External validity

Following the assumptions above, identifying the PATE involves a parallel set of assumptions for external validity:

**Conditional exchangeability for study selection**: $Y^a \perp S \,|X$ for all $a \in \mathcal{A}$. This assumption is also known as exchangeability over selection and the generalizability assumption. It requires that the outcomes among individuals with the same treatment and covariate values in the study and target populations are the same (Stuart *et al.*, 2011). All effect modifiers that differ between study and target populations must therefore be measured. This assumption would be satisfied by a study sample that is a random sample from the target population or a nonprobability study sample in which all effect modifiers are measured. A weaker condition, mean conditional exchangeability of selection, $E(Y^1 - Y^0|X, S = 1) = E(Y^1 - Y^0|X)$ can replace conditional exchangeability for study selection when focusing on the PATE (Kern *et al.*, 2016; Dahabreh *et al.*, 2019c).

**Positivity of selection**: $P(X = x) > 0 \Rightarrow P(S = 1|X = x) > 0$ with probability 1 for all $a \in \mathcal{A}$. This assumption requires common support with respect to study selection; in every stratum of effect modifiers, there is a positive probability of being in the study sample (Dahabreh *et al.*, 2017). This can be replaced by smoothing assumptions under a parametric model, for example, that the propensity score distribution has sufficient overlap or common support between the study sample and target population (Westreich *et al.*, 2017; Tipton *et al.*, 2017). Thus, with conditional positivity of selection we assume that all members of the target population are represented by individuals in the study. The positivity assumption in combination with the no unmeasured effect modification assumption above is also known as strongly ignorable sample selection given the observed covariates (Chan, 2017).

**SUTVA for study selection**: if $S = s$ (and $A = a$) then $Y = Y^a$. This assumption states that there is no interference between subjects selected into the study versus those not selected and that there is treatment version irrelevance between study and target samples (the same treatment is given to both) (Tipton, 2013a; Tipton *et al.*, 2017). It necessitates no difference across study and target samples in how outcomes are measured or in how the intervention is applied, that there is a common data-generating function for the outcome across individuals in the study and target populations (i.e., that being in the study does not change treatment effects), and that the potential outcomes are not a function of the proportion of individuals selected for the study. Treatment version irrelevance in SUTVA can be replaced by the condition of having the same distribution of treatment versions between study and target populations when estimating the PATE (Lesko *et al.*, 2017).

### 1.3.3 Transportability

Similar internal and external validity assumptions are needed for transportability, with the following modifications. When the study sample is a subset of the target population (generalizability), the positivity assumption for selection will need the propensity for selection to be bounded away from 0, whereas when the sample is not a subset of the target population (transportability), the propensity to be in the study population will need to be bounded

away from 0 and 1 (Tipton, 2013a). Furthermore, for transportability, the set of covariates, $X$, required for conditional exchangeability for study selection cannot include those that separate the study sample from the target population (e.g., hospital type if transporting results from teaching hospitals to community clinics, or geographic location if transporting between states) (Tipton, 2013a). Further distinctions are discussed by Pearl (2015) using the SEM framework. Under this framework, Pearl and Bareinboim formalize the assumptions necessary for using different transport formulas to reweight randomized data, providing graphical conditions for identifiability as well as transport formulas for randomized studies (Pearl and Bareinboim, 2014; Pearl, 2015), observational studies (Pearl and Bareinboim, 2011; Pearl, 2015; Bareinboim and Tian, 2015; Bareinboim and Pearl, 2016; Correa and Bareinboim, 2017; Correa *et al.*, 2018), and a combination of heterogeneous studies (Bareinboim and Pearl, 2014, 2016).

## 1.4 Assessing dissimilarity between target and study populations and testing for treatment effect heterogeneity

Numerous quantitative approaches can help evaluate the extent to which study results may be expected to generalize to the target population. These assessments examine population differences and whether treatment effect heterogeneity exists. Methods for assessing the similarity of study and target populations can broadly be categorized into those that compare baseline patient characteristics and those that compare outcomes for groups on the same treatment. For the former, many make use of the propensity score for selection, which also serves the purpose of assessing the extent to which propensity score adjustment using measured covariates can sufficiently remove baseline differences between study and target samples. However, most of these methods do not emphasize effect modifiers, hence should be combined with an assessment of whether the noted population differences correspond to heterogeneity of treatment effects. To test for heterogeneity of effects, one must first identify effect modifiers. Effect modifiers are often pre-specified by the investigator, but data-driven

approaches exist as well, and will be discussed in this section.

### 1.4.1 Assessing dissimilarity between populations using baseline characteristics

When summary-level study data are available, assessments that examine differences in univariate covariate metrics between study and target samples can be deployed. Cahan *et al.* (2017) propose a generalization score for evaluating clinical trials that incorporates baseline patient characteristics, the trial setting, protocol, and patient selection: it takes ratios of the mean or median values of these characteristics in the study and target samples, then averages across categories for an overall score. However, this approach does not account for any measures of dispersion, which may reflect exclusion of more heterogeneous individuals from the study. When only baseline patient characteristics are responsible for relevant study vs. target population differences, one can perform multiplicity-adjusted univariate tests for differences in effect modifiers between study and target samples (Greenhouse *et al.*, 2008). Alternatively, one could examine absolute standardized mean differences (SMD) for each covariate, $(\bar{X}_{\text{study}} - \bar{X})/\sigma_{\bar{X}}$, where $\bar{X}_{\text{study}}$ and $\bar{X}$ are the means of baseline covariates in the study and target samples, respectively, and $\sigma_{\bar{X}}$ is the standard deviation of $\bar{X}$ (Tipton *et al.*, 2017). High values indicate heavy extrapolation and reliance on correct model specification; in smaller samples, imbalances will often occur by chance (Tipton *et al.*, 2017). With one or more RCTs, generalizability across categorical eligibility criteria can be assessed by the percent of the target sample that would have been eligible for the study or set of studies (Weng *et al.*, 2014; He *et al.*, 2016; Sen *et al.*, 2016).

Joint distributions of patient characteristics can likewise be compared, such as by examining the SMD in propensity scores for selection (Stuart *et al.*, 2011). When the propensity score is not symmetrically distributed, summarizing mean differences is insufficient. Tipton (2014) developed a generalizability index that bins propensity scores and is bounded between 0 and 1: $\sum_{j=1}^{k} \sqrt{w_{p_j} w_{s_j}}$ with $j = 1,...,k$ bins, each with target sample proportions $w_{p_j}$ and study sample proportions $w_{s_j}$. It is based on the distributions of propensity scores rather than only the averages. However, this approach requires patient-level study and target sample

data. A generalizability index score of <0.5 indicates a study being very challenging to generalize from and a score of >0.9 indicates high generalizability (Tipton, 2014). Other propensity score distance measures can be used, such as Q-Q plots, Kolmogorov-Smirnov distance, Levy distance, the overlapping coefficient, and C statistic; these largely focus on comparing cumulative densities (Tipton, 2014; Ding *et al.*, 2016). To assess the degree of extrapolation with respect to effect modifiers, one can examine overlap in the propensity of selection distributions, such as the proportion of target sample individuals with propensity scores outside the $5^{th}$ and $95^{th}$ percentiles of the sample propensity scores (Tipton *et al.*, 2017).

One can also adopt a machine learning approach for detecting covariate shift–a change in the distribution of covariates between training and test data (here, the study and target data) (Glauner *et al.*, 2017). After creating a joint dataset with target and study sample data, a classification algorithm predicts whether the data came from the study. A dissimilarity metric surpassing a threshold of acceptability then indicates sizable dissimilarity between datasets. However, an inability to accurately predict study vs. target data origin does not rule out differences in effect modifiers. A low score might furthermore indicate an incorrect model specification or insufficient model tuning.

The tests discussed in this subsection assess differences between populations; however, they require investigator knowledge of which characteristics moderate the treatment effect (or are correlated with unmeasured effect modifiers) and what level of differences are clinically relevant. Many covariates are often tested or included in a propensity score regression for study selection. This approach prioritizes predictors that are strongly associated with study selection rather than those that exhibit strong effect modification. Investigators should therefore aim to identify relevant effect modifiers for testing or inclusion in the propensity score regression and test this subset.

### 1.4.2 Assessing dissimilarity between populations using outcomes

When individual-level outcome data or joint distributions of group-level outcome data are available in both the study and target samples for at least one of the treatment groups, the following methods can assess the extent to which measured effect modifiers account for population differences. One can compare the observed outcomes in the target sample to predicted outcomes using study controls (Stuart *et al.*, 2011), or more generally, study individuals who received the same treatment (Hotz *et al.*, 2005): $1/n_a \sum_{i=1}^{N} 1(A_i = a) Y_i$ vs. $1/n_{s,a} \sum_{i:S_i=1} 1(A_i = a) w_i Y_i$ with weights $w_i$ defined by weighting and matching methods discussed in Section 1.5.1. Hartman *et al.* (2015) formalize this comparison with equivalence tests. Alternatively, conditional outcomes for study and non-study target sample individuals receiving the same treatment, conditioning on measured effect modifiers, can be compared to detect unmeasured effect modification, although other identifiability assumption violations might also be at fault: $E(Y|X, A = a, S = 1)$ vs. $E(Y|X, A = a, S = 0)$. Possible tests include analysis of covariance, Mantel-Haenszel, U-statistic based tests, stratified log-rank, or stratified rank sum, depending on the outcome (Marcus, 1997; Hotz *et al.*, 2005; Luedtke *et al.*, 2019). For example, study controls could be compared to subgroups of the target population that were known to be excluded from the study (e.g., patients who declined participation in a RCT, as done by Davis (1988)). Relatedly, unmeasured effect modification can be imperfectly tested for by disaggregating a characteristic that differentiates the study from the target sample (Allcott and Mullainathan, 2012). These outcome differences should not exceed those observed between study treatment groups (Begg, 1992).

In addition to testing for outcome differences, one can test for differences between study and target regression coefficients or between baseline hazards in a Cox regression (Pan and Schaubel, 2009). Any identified differences in outcomes or effects will reflect sample differences unaccounted for by the outcome or weighting method, indicating unmeasured effect modification or an ineffective modeling approach. To have this comparison reflect relevant differences, study controls must be representative of the target population after weighting or regression adjustment. Hartman *et al.* (2015) provides a more formal set of

identifiability assumptions that may be violated when each equivalence test is rejected. If unmeasured effect modification is suspected, one can perform sensitivity analysis to assess the extent to which it can impact results (Marcus, 1997; Nguyen *et al.*, 2017, 2018; Dahabreh *et al.*, 2019d; Andrews and Oster, 2017) or to generate bounds on the treatment effect when only partial identification is possible (Chan, 2017).

### 1.4.3 Testing for treatment effect heterogeneity

Identified population differences are relevant insofar as they correspond to differences in treatment effect modifiers. The following tests enable an investigator to assess whether treatment effects vary substantially across measured covariates. Many are suitable for use in observational or RCT data, although have largely been demonstrated in RCT data to date. While some tests require a priori specification of subgroups, others can discover them in data-driven ways and most require individual-level data. A straightforward, but often overlooked issue is that studies with enrolled patients that are homogeneous with respect to effect modifiers will have difficulty identifying heterogeneity of effects. These approaches are therefore best applied to data representative of the target populations (Gunter *et al.*, 2011).

Tests of prespecified subgroups should focus on target population subgroups under- or over-represented in the study, or any other clinically relevant subgroup expected to exhibit effect heterogeneity. Largely, methods for testing treatment effect heterogeneity of a priori specified subgroups exhibit limited power. Those testing several effect modifiers individually are particularly underpowered to detect significant effects once multiple testing adjustments are incorporated. One approach tests the interaction term of treatment assignment with an effect modifier in a linear model, which also requires modeling assumptions as to the linearity and additivity of effects (Fang, 2017; Gabler *et al.*, 2009). To address this lack of power, sequential tests for identifying treatment-covariate interactions can be used with either randomized or observational data (Qian *et al.*, 2019). Alternative approaches, each addressing slightly different goals, include testing whether the conditional average treatment effect is identical across predefined subgroups (Crump *et al.*, 2008; Green and Kern, 2012),

comparing subgroup effects to average effects (Simon, 1982), and identifying qualitative interactions or treatment differences exceeding a prespecified clinically significant threshold (Gail and Simon, 1985).

When effect modifiers are not known a priori, a variety of techniques can be applied for identifying subgroups with heterogeneous effects. These include those that identify variables that qualitatively interact with treatment (i.e., for which the optimal treatment differs by subgroup) (Gunter *et al.*, 2011) as well as determine the magnitude of interaction (Chen *et al.*, 2017; Tian *et al.*, 2014). Various machine learning approaches can also be used to identify subgroups with heterogeneous treatment effects while minimizing modeling assumptions. Approaches that also present tests for treatment effect differences between subgroups include Bayesian additive regression trees (BART) and other classification and regression tree (CART) variants (Su *et al.*, 2008, 2009; Lipkovich *et al.*, 2011; Green and Kern, 2012; Athey and Imbens, 2016). Tree-based methods develop partitions in the covariate space recursively to grow toward terminal nodes with homogeneity for the outcome. These approaches may be particularly useful when heterogeneity may be a function of a more complex combination of factors.

With many effect modifiers or when effect modifiers are unknown, global tests for heterogeneity can also be used. Pearl (2015) provides conditions for identifying treatment effect heterogeneity (including heterogeneity due to unmeasured effect modifiers) for randomized trials with binary treatments, situations with no unobserved confounders, and with mediating instruments. Effect heterogeneity can be tested for using the baseline risk of the outcome as an effect modifier; interaction-based tests assess for differences in baseline risk between study and target population control groups (Varadhan *et al.*, 2016; Weiss *et al.*, 2012). These tests avoid the need for multiple testing but require outcome data in the target sample and modeling assumptions. A consistent nonparametric test also exists that assesses for constant conditional average treatment effects, $\tau_x = \tau \ \forall x \in \mathcal{X}$ (Crump *et al.*, 2008). Additional methods, which suffer from limited power and rely on estimates of SATE, include testing whether potential outcomes across treatment groups have equal variances and whether cumulative

distribution functions of treatment and control outcomes differ by a constant shift (Fang, 2017). Global tests do not identify subgroups responsible for effect heterogeneity, although if a global test is rejected, one can then compare individual subgroups to determine which demonstrate effect heterogeneity.

If these assessments of generalizability fail and the target population is not well-represented by the study population (specifically, when strong ignorability fails), Tipton (2013a) provides several recommended paths forward. Investigators can change the target population to one represented by the study. That is, change the estimand of interest by aligning inclusion and exclusion criteria, outcome timepoints, or treatment doses (Hernán *et al.*, 2008; Weisberg *et al.*, 2009). A population coverage percentage can then summarize the percent overlap between the new and original target sample propensity scores, and describe relevant differences from the original target population. Investigators can alternatively retain the original target population and note the limitations of extrapolated results and likelihood of remnant bias. It is also important to acknowledge that a different study may need to be conducted.

## 1.5 Generalizability and transportability methods for estimating population average treatment effects

Following the application of the methods in the previous sections, including assessing the plausibility of relevant assumptions, an analytic method is typically needed to generalize or transport results from randomized or observational data to a target population. These approaches have many parallels to those used to address internal validity bias. We revisit weighting and matching-based methods and outcome regressions in depth while additionally examining techniques that use both propensity and outcome regressions (these are often doubly robust). To mitigate external validity bias, generalizability and transportability methods address differences in the distribution of effect modifiers between study and target populations. To do so, for weighting and matching-based approaches, these methods account for the probability of selection into the study, rather than the probability of treatment

assignment. Outcome regressions require that treatment effect is allowed to vary across all effect modifiers in addition to all confounders being correctly included in the regression.

Most generalizability and transportability methods have been developed for randomized data. When outcome data are available from both randomized studies and an observational study representative of the target population, their combination has the potential to overcome sensitivity to positivity violations for selection into the study (an issue that RCT data commonly face) as well as to unmeasured confounding (which may afflict observational studies). Incorporating observational data in a principled manner can also shrink mean squared error. However, many such approaches do not leverage the internal validity of RCT data. The following sections will highlight some exceptions. While most approaches require individual-level study and target sample data, Appendix A highlights approaches that only use summary-level data for either the study or target sample.

### 1.5.1 Weighting and matching methods

Methods that adjust for differing baseline covariate distributions between study and target samples via weighting or matching are particularly effective when effect modifiers strongly predict selection into the study. While including unnecessary covariates can decrease precision, increase the chance of extreme weights and difficult-to-match subjects, and provide no bias reduction (Nie *et al.*, 2013), failing to include an effect modifier is typically of greater concern than including unnecessary covariates (Stuart, 2010; Dahabreh *et al.*, 2018). Matching and reweighting methods strongly rely on common covariate support between study and target populations and perform poorly when a portion of the target population is not well-represented in the study sample or when empirical positivity violations occur. Investigators should use the estimation approach that leads to the best effect modifier balance for their study (Stuart, 2010) and strive for fewer assumptions.

### Matching

Full matching and fine balance of covariate first moments (i.e., expected values) have been used in the generalizability context (Stuart *et al.*, 2011; Bennett *et al.*, 2020). Stuart *et al.* (2011) fully match study and target sample individuals based on their propensity scores to form sets so that each matched set has at least one study and target individual. Individuals' outcomes are then reweighting by the number of target sample individuals in their matched set. This approach relies heavily on the distance metric used, which can be misled by covariates that don't affect the outcome. Fine balance of covariate first moments is a nonparametric approach for larger data that can also be used with multi-valued treatments (Bennett *et al.*, 2020). This approach matches samples to a target population to achieve fine balance on the first moments of all covariates rather than working with the propensity score.

Some implementations of these methods only match a subset of study individuals (hence show areas of the covariate distribution without common support), while others ensure all study and target sample individuals are matched. Matching methods require calibration for bias-variance tradeoff such as via a caliper or by choosing the ratio of study to target individuals to match. A variety of distance metrics exist; however, none specifically target effect modifiers. With unrepresentative observational data, treatment groups can first be matched based on confounding variables before matching study pairs to the target sample based on effect modifiers, or each treatment group can be separately matched to the target sample (Bennett *et al.*, 2020).

### Weighting

**Post-stratification.** In a low-dimensional setting with categorical or binary covariates, one can use nonparametric post-stratification (also known as direct adjustment or subclassification), as has been done in the literature with randomized data (Miettinen, 1972; Prentice *et al.*, 2005) and with observational data in the context of instrumental variables (Angrist and Fernández-Val, 2013). Post-stratification consists of obtaining estimates for each stratum of effect modifiers, then reweighting these estimates to reflect the effect modifier distribution in

the target population, i.e., $\hat{E}(Y^a) = 1/n \sum_{l=1}^{L} n_l \bar{Y}_l^a$, where $L$ is the number of strata, $n_l$ is the target sample size in stratum $l$, $n = \sum_{l=1}^{L} n_l$, and $\bar{Y}_l^a$ is an estimate from study sample data of the potential outcome on treatment $a$ in stratum $l$, commonly the stratum-specific sample mean for subjects on treatment $a$ (Miettinen, 1972; Prentice *et al.*, 2005).

Post-stratification only requires stratum-specific summary data and closed-form variance formulas are often available. However, empty strata quickly become an issue when dealing with continuous variables or many stratifying variables. Conversely, if insufficient strata are used, residual external validity bias will remain, which is particularly problematic in small samples (Tipton *et al.*, 2017). To combat this, inference can be pooled across strata using multilevel regression with post-stratification (Pool *et al.*, 1964; Gelman and Little, 1997; Park *et al.*, 2004; Kennedy and Gelman, 2019).

For higher dimensional settings or with continuous covariates, more flexible nonparametric approaches can be applied, such as maximum entropy weighting, where study strata are reweighted to the distribution in the target sample (Hartman *et al.*, 2015). When target and study populations differ on post-treatment variables such as adherence, principal stratification can be used to estimate PATEs by classifying subjects into never-taker, always-taker, and complier categories (Frangakis, 2009).

**Estimating using the propensity for study selection.** Most weighting approaches use a propensity of selection regression to construct weights. They rely on correct specification of the propensity score regression and sufficient overlap in propensity scores between study subjects and target sample individuals not in the study. These approaches have the additional advantage of allowing one set of weights to be used for treatment effects related to multiple outcomes. The most straightforward weighting approaches tend to have large variances in the presence of extreme weights, give disproportionate weight to outlier observations, and produce outcome estimates outside the support of the outcome variable. Weight standardization can address these issues, as can weight trimming, although the latter induces bias by changing the target population of interest, hence requiring a careful bias-variance trade-off.

Inverse probability of participation weighting (IPPW), a Horvitz-Thompson-like approach

(Horvitz and Thompson, 1952), is the most common weighting technique for generalizability (Flores and Mitnik, 2013; Baker *et al.*, 2013; Lesko *et al.*, 2017; Westreich *et al.*, 2017; Correa *et al.*, 2018; Dahabreh *et al.*, 2018, 2019c). Most simply, IPPW weights the outcome for each study individual on treatment $a$ by the inverse probability (propensity) of being in the study. Weights have been developed for estimating PATEs, including those that incorporate treatment assignment to account for covariate imbalances in an RCT or for confounding in an observational study. The observed outcomes are reweighted to obtain the potential outcomes for each treatment group $a$: $E(Y^a) = \frac{1}{n}\sum_{i=1}^{n} w_i Y_i$ with

$$w_i = \frac{1}{\pi_{s,i}} I(S_i = 1) I(A_i = a) \qquad \text{for random treatment assignment (Lesko } \textit{et al.}\text{, 2017)}$$

$$w_i = \frac{1}{\pi_{s,i}\pi_{a,i}} I(S_i = 1) I(A_i = a) \qquad \text{more generally (Stuart } \textit{et al.}\text{, 2011; Dahabreh } \textit{et al.}\text{, 2019c),}$$

where $I(S_i = 1)$ is the indicator for being in the study, $I(A_i = a)$ is the indicator for being assigned treatment $a$, $\pi_{s,i} = P(S_i = 1|X_i)$ is the propensity score for selection into the study and $\pi_{a,i} = P(A_i = a|S_i = 1, X_i)$ is the propensity score for assignment to treatment $a$ in the study.

Individual-level data are typically required, although one can also use joint covariate distributions from group-level data (Cole and Stuart, 2010) or univariate moments (e.g., means, variances) with additional assumptions (Signorovitch *et al.*, 2010; Phillippo *et al.*, 2018). Because IPPW only uses study individuals on a given treatment to estimate potential outcomes for that treatment, power can become an issue, particularly for multi-level treatments. These methods also perform poorly when study selection probabilities are small, which can be a common occurrence for generalizability (Tipton, 2013a). IPPW weights have also been developed for regression parameters in a generalized linear model (Haneuse *et al.*, 2009), as well as for Cox model hazard ratios and baseline risks (Cole and Stuart, 2010; Pan and Schaubel, 2008).

For transportability to the target population $S = 0$, odds of participation weights are used rather than inverse probability of participation weights (Westreich *et al.*, 2017; Dahabreh *et al.*, 2018). This corresponds to the estimator $E(Y^a|S = 0) = \frac{1}{n}\sum_{i=1}^{N} w_i Y_i$ with $N = n + n_s$

and weights (Dahabreh *et al.*, 2018):

$$w_i = \frac{1 - \pi_{s,i}}{\pi_{s,i}\pi_{a,i}} I(S_i = 1) I(A_i = a).$$

To address potentially unbounded outcome estimates, standardization then replaces $n$ by the sum of the weights, which normalizes the weights to sum to 1 (Dahabreh *et al.*, 2018, 2019c). The resulting estimator will be more stable, bounded by the range of the observed outcomes, and perform better when the target sample is much larger than the study.

Under regularity conditions, estimates derived using IPPW are consistent and asymptotically normal (Lunceford and Davidian, 2004; Pan and Schaubel, 2008; Cole and Stuart, 2010; Correa *et al.*, 2018; Buchanan *et al.*, 2018). Variance for the IPPW estimator can be obtained through either a bootstrap approach or robust sandwich estimators. The latter may be difficult to calculate (Haneuse *et al.*, 2009) and bootstrap methods for IPPW have been shown to perform better when there is substantial treatment effect heterogeneity or smaller sample sizes (Chen and Kaizar, 2017; Tipton *et al.*, 2017).

Propensity scores can also be used in the context of post-stratification, weighting or matching individuals within strata. RCT individuals are divided into strata defined by their propensity scores; quintiles are commonly used, based on results showing that this approach may remove over 90% of bias (O' Muircheartaigh and Hedges, 2014). Effects are estimated using sample data within each subgroup, such as through separate regressions or a joint parametric regression with fixed effects for subgroups and interaction terms for subgroups by RCT status. Results can then be reweighted based on the number of target sample individuals in each subgroup (O' Muircheartaigh and Hedges, 2014). Alternatively, the target sample can be matched to RCT individuals within the same propensity score stratum (Tipton, 2013a).

The post-stratification estimator is asymptotically normal and closed-form variance estimates exist for independent strata (O' Muircheartaigh and Hedges, 2014; Lunceford and Davidian, 2004). Compared to IPPW, strata reweighting is more likely to be numerically stable and easily implementable when treatment assignment is done at the group level (e.g., cluster-randomized trials). However, stratification implicitly assumes that treatment effects

are identical for study and target patients in the same stratum; this assumption is rarely met, resulting in residual confounding and inconsistent estimates (Lunceford and Davidian, 2004). It also relies on the assumptions of treatment effect heterogeneity being fully captured by the propensity score for treatment and that outcomes are continuous and bounded. With too few strata, bias reduction will be insufficient; conversely, too many strata can lead to small strata counts and unstable estimates (Stuart, 2010; Tipton *et al.*, 2017).

Propensity strata approaches have also been used to address positivity of treatment assignment violations within the target sample in the setting where outcome data are available from both a randomized and observational study (Rosenman *et al.*, 2018). Rosenman *et al.* (2020) present an extension which aims to adjust for potential unmeasured confounding bias.

### 1.5.2 Outcome regression methods

#### Outcome data from one study

Outcome regressions, also known as response surface modeling, have not been as extensively developed for generalizability and transportability compared to propensity-based approaches. Broadly speaking, outcome regressions approaches fit an outcome regression in study sample data to estimate conditional means, then obtain PATEs by marginalizing over (i.e., standardizing to) the target sample covariate distribution by predicting counterfactuals for the target sample: $\hat{E}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} \hat{E}(Y_i|S_i = 1, A_i = a, X_i)$. If the target sample is not a simple random sample from the target population, this would be a weighted average using sampling weights (Kim *et al.*, 2018).

Outcome regression approaches are particularly effective when effect modifiers strongly predict the outcome and when the outcome is common but selection into the study is rare. They are also convenient for exploring PCATEs. These approaches can yield better precision than weighting or matching-based methods because they can adjust both for confounders, effect-modifiers, and factors only predictive of the outcome, thus decreasing variance in the estimate. They are simple to implement when an outcome regression for confounding

adjustment has already been fit and accounts for all relevant effect modifiers. The same regression that was used to estimate impacts within the study can then be used to predict counterfactuals in the target sample. Outcome regression methods can be used with either randomized or observational study data, but have been used most frequently in RCTs. In the presence of significant non-overlap between the target and study samples, outcome regressions rely on heavy extrapolation (Kern *et al.*, 2016; Attanasio *et al.*, 2003), often with no corresponding inflation of the variance to reflect uncertainty in the resulting estimates.

The simplest approach is an ordinary least squares outcome regression (Flores and Mitnik, 2013; Kern *et al.*, 2016; Elliott and Valliant, 2017; Dahabreh *et al.*, 2018, 2019c). An outcome regression is fit with interaction terms between treatment and all effect modifiers before predicting counterfactual outcomes for the target sample (the marginalization step). Dahabreh *et al.* (2018) showed the consistency of this type of outcome regression for the PATE. For RCTs, separate regressions are recommended for each treatment group to better capture treatment effect heterogeneity (Dahabreh *et al.*, 2019c), although this approach precludes borrowing information across treatment groups, which is possible with machine learning methods that discover treatment effect heterogeneity.

Among these machine learning techniques is BART, which is the most commonly used data-adaptive outcome regression approach for generalizability and transportability (Chipman *et al.*, 2007, 2010; Kern *et al.*, 2016; Hill, 2011). Tree-based methods, including BART, were briefly introduced in Section 1.4.3. BART models the outcome as a sum of trees with linear additive terms and a regularization prior. BART addresses external validity bias via its data-driven discovery of treatment effect heterogeneity and strengths of the method include its ability to obtain confidence intervals from the posterior distribution (Hill, 2011; Green and Kern, 2012). However, BART credible intervals show undercoverage when the target population differs substantially from the RCT (Hill, 2011).

Data availability may challenge these outcome regression approaches. When the covariates in the target sample aren't available in the study sample, or vice versa, but the SATE can be expected to be approximately unbiased for the PATE, the SATE estimates' credible

intervals can be expanded to account for uncertainty in the target population covariate distribution (Hill, 2011).

**Outcome data from multiple studies**

Here, we consider meta-analytic approaches for summary-level data as well as studies that combine individual-level data from more than one study (for example, one randomized and one observational study). Much of the literature has focused on meta-analytic techniques using summary-level study data and no target sample covariate information. This body of bias-adjusted meta-analysis methods largely does not explicitly define a target population for whom inference is desired, but rather relies on subjective investigator judgments of the levels of bias in each study, specified using bias functions or priors in a Bayesian framework. Eddy (1989) presents the first such approach, the confidence profile method for combining chains of evidence. Likelihoods are adjusted for different study designs' (investigator-specified) internal and external validity biases; uncertainty around these biases are incorporated through prior distributions. Various subsequent Bayesian hierarchical models have been developed, such as a 3-level model (Prevost *et al.*, 2000) with the levels corresponding to models of the observed evidence, variability between studies, and variability between study types (randomized vs. observational). When available, covariate information can be added to the models to address effect heterogeneity. Effectively, this estimator averages across the internal and external validity biases of the studies and therefore is only unbiased when the external validity bias in the RCT exactly 'cancels' the internal validity bias in the observational data (Kaizar, 2011).

Other meta-analysis studies leveraging summary-level data separately specify internal and external validity bias parameters for an explicit target population and down-weight studies with higher risk of bias. One such example is the bias adjusted meta-analysis approach by Turner *et al.* (2009), which presents a checklist that subjectively quantifies the extent of internal and external validity bias for each study and then weighs studies' average outcomes by the extent of bias. Greenland (2005) pool across observational case-control studies using a

Bayesian meta-sensitivity model with bias parameters to separately permit consideration of misclassification, non-response, and unmeasured confounding. In the intermediate setting where individual-level data is available in the study but only covariate moments (e.g., means, variances) are available in the target setting, Phillippo *et al.* (2018) present an outcome regression approach for indirect treatment comparison across RCTs.

When individual-level outcome data is available in the target sample or from multiple studies, data can be combined into one joint dataset for outcome regression analysis if the outcome regression can be expected to be the same across studies (Kern *et al.*, 2016). Such an approach can be preferential to IPPW, which uses only study and not target sample outcome data (Kern *et al.*, 2016). However, it will be dominated by observational data results (and their potential biases) when observational subjects constitute the majority of the joint dataset, effectively result in a weighted average across studies, weighted by the proportion of subjects in each study.

Hierarchical Bayesian evidence synthesis is the only outcome regression approach we identified that attempts to empirically adjust for unobserved confounding when estimating effects for observational patients who are not well-represented in the RCTs (Verde *et al.*, 2016; Verde, 2019). Summary-level RCT data are combined with individual-level observational data through a weighting approach in which the control group event rate is assumed to be similar across all studies and a study quality bias term is added to the observational studies' outcome regression to account for unmeasured confounding or other uncontrolled biases and to inflate variance. Alternatively, Gechter (2015) derive bounds on the PATE and PCATE when transporting from an RCT to a target sample with outcome data (all untreated).

### 1.5.3 Combined propensity score and outcome regression methods

#### Outcome data from one study

Double robust methods for generalizability and transportability typically combine outcome and propensity of selection regressions. They are asymptotically unbiased when at least one of these regression functions is consistently estimated, and if both are consistently es-

timated, asymptotically efficient. However, if neither regression is estimated consistently, the mean squared error may be worse than using a propensity or outcome regression alone. Incorporation of flexible modeling approaches can help mitigate regression misspecification. Three asymptotically locally efficient double robust approaches have been developed in randomized data: a targeted maximum likelihood estimator (TMLE) for instrumental variables (Rudolph and van Der Laan, 2017), which is a semiparametric substitution estimator, the estimating equation-based augmented inverse probability of participation weighting (A-IPPW) (Dahabreh *et al.*, 2018, 2019c), and an augmented calibration weighting estimator that can also incorporate outcome information from the target sample when it is available (Dong *et al.*, 2020).

The TMLE was developed for transportability in an encouragement design setting (i.e., intervention focused on encouraging individuals in the treatment group to participate in the intervention) with instrumental variables (Rudolph and van Der Laan, 2017) and has also been used for generalizability (Schmid *et al.*, 2020). Three different PATE estimators were developed: intent to treat, complier, and as-treated. All use an outcome regression to obtain an initial estimate, then adjust that estimate with a fluctuation function using a clever covariate $C$, which is derived from the efficient influence curve and incorporates the propensity of selection information in a bias reduction step. For example, for the intent to treat PATE, the fluctuation function takes the form: $\text{logit}(\hat{E}(Y|S=1,A,Z,X)+\epsilon C)$, where

$$C = \frac{I(S=1,A=a)}{P(A=a|S=1,X)P(S=1)} \frac{P(Z=z|S=0,A=a,X)P(X|S=0)}{P(Z=z|S=1,A=a,X)P(X|S=1)}$$

and $Z$ corresponds to the intervention taken (whereas $A$ corresponds to the assigned intervention, as before). The approach allows outcome and propensity regressions to be flexibly fit, for example, using an ensemble of machine learning algorithms. Variances are calculated from the influence curve.

A-IPPW has been developed both for generalizing results to estimate PATEs for all trial-eligible individuals (Dahabreh *et al.*, 2019c,a) and for transporting results to estimate PATEs for trial-eligible individuals not included in a trial (Dahabreh *et al.*, 2018). Three

double robust estimating equation-based estimators are presented: A-IPPW, A-IPPW with normalized weights that sum to 1 to ensure bounded estimates, and a weighted outcome regression estimator using participation weights. The non-normalized A-IPPW estimators are as follows, with $w_i$ the same as for IPPW:

$$\frac{1}{n}\sum_{i=1}^{n}\{w_i\{Y_i-\hat{E}(Y_i|S_i=1,A_i=a,X_i)\}+\hat{E}(Y_i|S_i=1,A_i=a,X_i)\} \quad \text{for generalizability}$$

$$\frac{1}{n}\sum_{i=1}^{N}\{w_i\{Y_i-\hat{E}(Y_i|S_i=1,A_i=a,X_i)\}+\{1-I(S_i=1)\}\hat{E}(Y_i|S_i=1,A_i=a,X_i)\} \quad \text{for transportability}$$

Variance can be derived using empirical sandwich estimates or using a nonparametric bootstrap. As these estimators are partial M-estimators, they can produce estimates outside bounds if the outcome regression is not well-chosen and they may have multiple solutions.

Several other double robust estimators for transportability resemble the IPPW estimator, with sampling weights derived through alternative approaches that do not rely on propensity scores (Josey *et al.*, 2020b,a; Dong *et al.*, 2020). For example, the semiparametric and efficient augmented weighting estimator by Dong *et al.* (2020) calibrates the RCT covariate distribution to match that of the sampling-weighted target sample.

An alternative reweighted outcome regression method for observational data does not claim double robustness and draws from the unsupervised domain adaptation literature. In general, unsupervised domain adaptation methods aim to make predictions for a target sample (the "target domain") when outcomes are only observed in the study sample ("source domain"). The approach of Johansson *et al.* (2018) is a regularized neural network estimator for PCATE parameters that jointly learns representations from the data and a reweighting function. Representational learning creates balance between the study and target covariate distributions and between treated and control distributions in a representational space so that predictors use information common across these distributions and focus on covariates predictive of the outcome. In this learned representational space, results are then re-weighted to minimize an upper bound on the expected value of the loss function under the target covariate distribution. Propensity scores can also be used to reweight a likelihood function, as done by Nie *et al.* (2013) in an RCT setting for calibrating control outcomes from prior

studies to the trial target sample. Similarly, Flores and Mitnik (2013) reweight an outcome regression to the target sample.

**Outcome data from multiple studies**

Several methods have combined randomized and observational data sources such that that they retain the internal validity of the randomized data and the external validity of the target sample observational data. These approaches broadly rely on the assumption that the relationship between unmeasured confounders and potential outcomes is the same in the RCT as in the target sample, which is a weaker assumption than that of no unmeasured confounding required by most of the methods described thus far. One study combined individual-level data from several RCTs to transport results to the target sample, extending the A-IPPW estimator (as well as corresponding IPPW and outcome regression estimators) to the multi-study setting (Dahabreh *et al.*, 2019b). The remainder of the section discusses approaches that combine randomized and observational data.

When differences in effect modifiers between the RCT and target population are known (e.g., by inclusion and exclusion criteria), cross-design synthesis meta-analysis is a method for combining randomized and observational study data while capitalizing on the internal validity of the randomized data and the external validity of the observational data (Begg, 1992; Greenhouse *et al.*, 2017). It provides a means for estimating treatment effects for patients excluded from the RCT and can use summary-level RCT data if outcomes are available by relevant patient subgroups, although can only accommodate a limited number of strata of relevant effect modifiers.

Cross-design synthesis meta-analysis effectively assumes a constant amount of unmeasured confounding across patients eligible and ineligible for the RCTs (Kaizar, 2011). This approach will have smaller bias than use of randomized or observational data alone under various common data scenarios and, across simulations, shows better coverage through smaller bias and increased variance (Kaizar, 2011).

When differences between RCT and target populations are less well understood, there

are continuous effect modifiers, or a higher dimensional set of effect modifiers, one can use Bayesian calibrated risk-adjusted regressions (Varadhan *et al.*, 2016; Henderson *et al.*, 2017). This parametric approach requires individual-level information from observational and randomized studies, leveraging outcome regressions and calibration using the propensity of selection. The target population is assumed to be represented by a subset of the observational data; the RCT data are likewise assumed to be represented by a (potentially different) subset of the observational data. The calibrated risk-adjusted model performs well when there is poor overlap between RCT and target data; however, it relies on the observational dataset having substantial effect modifier overlap with both the target sample and RCT. Robust variance formulas or bootstrapping can be used to obtain confidence intervals.

A 2-step frequentist approach for consistently estimating PCATE parameters has been developed to estimate effects in a target population represented by observational data (Kallus *et al.*, 2018). It begins with outcome regressions for each treatment group of the observational data, or a flexible regression that captures effect heterogeneity. Observational data are then standardized to the RCT population before 'debiasing' their estimates using RCT data by including a correction term that can depend on measured covariates. This method relies on the assumption that calibrating internal validity bias in the subset of the observational data distribution overlapping with RCT data appropriately calibrates the bias for the entire target sample. The 2-step approach would therefore not necessarily decrease bias if the covariate distribution is highly imbalanced, resulting in average biases that are quite different between the RCT overlapping vs. nonoverlapping subsets of the target sample.

Lu *et al.* (2019) present an approach that, unlike the above methods, assumes no unmeasured confounding in the observational data when combining RCT and comprehensive cohort study data (where patients who decline randomization are enrolled in a parallel observational study). They use semiparametric double robust estimators that can incorporate flexible regressions.

## 1.6 Discussion

Obtaining unbiased estimates for a relevant target population requires applying generalizability or transportability methods in studies that meet required identifiability assumptions. The internal validity of randomized trials is not sufficient to obtain unbiased causal effects; external validity also needs to be considered. In this synthesis, we have discussed (1) sources of external validity bias and study designs to address it, (2) defining an estimand in a target population of interest, (3) the identifiability assumptions underpinning generalizability and transportability approaches, (4) a variety of approaches for quantifying the relevant dissimilarity between study and target samples and assessing treatment effect heterogeneity, and (5) a variety of matching and weighting methods, outcome regression approaches, and techniques that use both outcome and propensity regressions that generalize or transport from randomized and observational studies to a target population. These approaches have been applied across diverse settings from RCT results transported to patients represented in registries to cluster-randomized educational intervention trials generalized to broader geographic areas. Across a variety of settings, it is important to estimate results for populations that go beyond the study population. We suggest the following considerations for researchers.

*Make efforts to explicitly define the target population(s) and identify the study population from which your study sample data is a simple random sample.* Describing the study population may be a difficult task, and there may not be a practically meaningful population that is representative of your study sample data. However, this clarity will allow you to compare and, when feasible, better-align the study sample data to the target population. Discussion regarding target population(s) should be guided by the ensuing decisions the study aims to inform as well as practical considerations (e.g., lack of certain subgroups in your study). These considerations may require iteration between feasibility and the desired study aims as well as careful discussion amidst study collaborators. When combining across studies, meta-analyses should likewise carefully specify target population(s) for inference and incorporate considerations of treatment effect heterogeneity or demonstrate that effect heterogeneity is

not a concern. Without transparency in the target population(s), a study cannot estimate well-defined treatment effects nor can readers judge the generalizability of study results to any other population of interest.

*Plan for generalization in your study design, when feasible, including writing generalizability considerations into your grant or study objectives.* Enroll randomized study participants or design observational study inclusion and exclusion criteria to have the study sample be representative of the target population, or fully capturing the heterogeneity of effect modifiers. Collect data on likely treatment effect modifiers that are associated with study participation. Attempt to identify and mitigate potential sources of missingness or selection bias. If possible, collect baseline characteristics and outcome data on study nonparticipants who are part of the target population. Otherwise, identify external sources of data that might inform the composition of your target population with respect to effect modifiers and work towards aligning variables between these target sample data sources and your study.

*Clearly describe the internal and external validity assumptions needed to identify the treatment effect as they relate to your study.* Substantively assess the justifiability of these internal and external validity assumptions. To the extent possible, test the validity of the assumptions and perform sensitivity analyses to assess the impact of assumption violations.

*Quantify the dissimilarity between the study and target populations using at least one method.* Ideally, use multiple methods, as they each tell different parts of the story: examine univariate and joint distributions of effect modifiers, differences in the propensity to participate in the study, and (if outcome information is available in the target sample) differences in outcomes between study and target subjects on the same treatment. If differences are identified, one should investigate which subpopulations drive those differences and assess whether they have heterogeneous treatment effects. In addition to examining subject characteristics, assess whether differences exist in the setting, treatment, or outcome.

*To obtain causal estimates when the target and study populations differ with respect to effect modifiers, incorporate at least one generalizability or transportability estimator.* Alternatively, at the minimum, assess and describe sources of effect heterogeneity and whether they're

likely to differ for the target population. Derive estimates using as much data as possible (e.g., when outcome data is available, use it in a principled way). The choice of method for external validity bias adjustment may be restricted by data availability (e.g., summary-level vs. individual-level data) but should be driven by similar principles as those that guide the choice between outcome regressions, matching and weighting methods, and double robust approaches for confounding adjustment (Van der Laan *et al.*, 2003; Neugebauer and van der Laan, 2005; van der Laan and Rose, 2011). Flexible nonparametric and semiparametric models and estimators that use ensemble machine learning minimize the need for strict parametric assumptions and have the potential to perform the best (Kern *et al.*, 2016).

*For both methods developers and applied researchers, we recommend releasing publicly available code alongside the paper and providing details for implementation.* Published code facilitates replicability and accessibility of methods for future research and applied use. A substantial barrier to the adoption of new statistical methods, including advances in generalizability and transportability, is the lack of available computational tools.

While much of the causal inference literature has focused on issues of internal validity, both internal and external validity are necessary for valid inference. When treatment effect heterogeneity exists, as is often the case, study results may not hold for a target population of interest. Approaches to address internal validity biases can be borrowed to improve upon methods for addressing external validity bias. This review presents a framework for such analysis and summarizes different choices for estimators that can be used to generalize or transport results to a population different from the one under study. It brings together diverse cross-disciplinary literature to provide guidance both for applied and methods researchers. Improving the incorporation of results from observational studies, including electronic health databases, can lead to better inference for policy-relevant populations with reduced bias and improved precision.

# Chapter 2

# Conditional Cross-Design Synthesis Estimators for Generalizability in Medicaid

Irina Degtiar[1], Tim Layton[2], Jacob Wallace[3], Sherri Rose[4]

[1] *Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA*

[2] *Department of Health Care Policy, Harvard Medical School, Boston, MA, USA*

[3] *Department of Health Policy & Management, Yale School of Public Health, New Haven, CT, USA*

[4] *Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, CA, USA*

## Abstract

While much of the causal inference literature has focused on addressing internal validity biases, both internal and external validity are necessary for unbiased estimates in a target population of interest. When the target population is not well-represented by a randomized study, but is reflected when incorporating observational data, few generalizability approaches exist to estimate causal quantities in the target population. We propose a class of novel conditional cross-design synthesis estimators that combine randomized and observational data, while addressing the respective biases of these data sources, to generalize to a target population represented by a union of the data. The estimators include outcome regression, propensity weighting, and double robust approaches. All use the covariate overlap between the randomized and observational data to remove potential unmeasured confounding bias. We apply these methods to estimate the causal effect of managed care plans on health care spending among New York City Medicaid beneficiaries.

## 2.1 Background

When estimating causal effects, randomized data estimates often have internal validity—unbiased causal effects for the population represented by the study. However, these estimates may not reflect causal effects in the target population (i.e., external validity), and, furthermore, not represent subsets of the target population. Observational data may be more representative of the target population and, hence, have external validity, but are potentially affected by unmeasured confounding. These challenges arise in settings ranging from clinical trials that exclude certain patient subsets (Prentice *et al.*, 2005; Lu *et al.*, 2019) to policy evaluation studies that aim to inform deployment in a different population (Attanasio *et al.*, 2003; Kern *et al.*, 2016). While much of the causal inference literature has focused on addressing internal validity biases, both internal and external validity are necessary for unbiased estimates.

Although generalizability and transportability methods exist for extending inference from a randomized study to a target population, few leverage a combination of randomized and observational data to address each data source's shortcomings (Degtiar and Rose, 2021). Approaches that do combine individual-level randomized and observational data face limitations when the target population doesn't fully overlap with the randomized data and the observational data have unmeasured confounding. Existing techniques extrapolate from the randomized data beyond their support (Attanasio *et al.*, 2003; Hill, 2011; Kern *et al.*, 2016), assume the included observational data have no unmeasured confounding (Kern *et al.*, 2016; Lu *et al.*, 2019), or allow for unmeasured confounding but assume treatment effects are identical within strata of effect modifiers, which may not hold with continuous effect modifiers (Rosenman *et al.*, 2020). Cross-design synthesis methods combine randomized and observation data, often relying on a binary flag that determines eligibility in the randomized subset of the data, which requires overlap membership to be known (Begg, 1992; Kaizar, 2011; Greenhouse *et al.*, 2017). Bayesian calibrated risk-adjusted modeling, currently only deployed in the context of Cox proportional hazards survival regression, necessitates a third external data source that has strong overlap with both the randomized and observational

data (Varadhan *et al.*, 2016; Henderson *et al.*, 2017). A 2-step regression approach by Kallus *et al.* (2018) assumes that the randomized covariate distribution is subsumed in the observational data distribution. The method also does not directly extend to estimating population treatment-specific means rather than average treatment effects.

We present a novel class of methods, which we refer to as conditional cross-design synthesis (CCDS) estimators, addressing several limitations of existing estimators that incorporate outcome information from randomized and observational data. All CCDS approaches estimate a conditional bias term from the overlapping support between randomized and observational data that is then used to 'debias' observational data estimates. These techniques are robust to unmeasured confounding in the observational data and positivity violations for selection into the randomized data. The estimators include outcome regression, 2-step outcome regression, inverse probability weighting, and double robust augmented inverse probability weighting approaches. Our implementation allows for the incorporation of ensemble machine learning to estimate the regression components of the various estimators, minimizing reliance on misspecified parametric regressions.

We apply our class of CCDS estimators to a study in New York City (NYC) Medicaid Managed Care (MMC), which provides health insurance to most New York Medicaid beneficiaries. Beneficiaries who do not choose a health plan are randomly assigned to one. However, the 7% of NYC beneficiaries who are randomized are not representative of the broader NYC Medicaid population. Of the remaining 93% of enrollees who actively chose their health plan (i.e., the observational beneficiaries), some are not well-represented by any randomized beneficiaries. This motivates our CCDS approaches that combine randomized and observational data to estimate health plan-specific causal effects on health care spending in the full NYC Medicaid population.

Section 2.2 defines notation and the estimand of interest. Section 2.3 reviews standard generalizability assumptions, describes our relaxation of two of the assumptions through the combination of randomized and observational data, and identifies the estimand of interest under our relaxed assumptions. Section 2.4 presents the novel CCDS estimators and the

limited available alternative approaches. We evaluate all estimators through a simulation study in Section 2.5 that highlights settings where each CCDS estimator can be anticipated to perform well. Section 2.6 applies these methods to our NYC Medicaid study examining the impact of managed care plans on health care spending. Section 2.7 concludes with a discussion.

## 2.2 Notation and estimand

### 2.2.1 Notation

The target population of interest is represented by a target sample. A portion of the target sample is randomized to the intervention (i.e., managed care plans) and the remaining individuals are observational. Hence the target sample is a union of randomized and observational data. We observe $n = n_{\text{RCT}} + n_{\text{obs}}$ independent draws from an underlying probability distribution $P \in \mathcal{M}$, where $\mathcal{M}$ is statistical model, namely, a collection of possible probability distributions. Each of these draws consist of an outcome $Y \in \mathbb{R}$, the intervention $A \in \mathcal{A}$, the vector of covariates $\boldsymbol{X} \in \mathcal{R} \in \mathbb{R}^d$, where $\mathcal{R}$ is the region of support in the target population's covariate distribution, and an indicator for selection into the randomized group $S \in \mathcal{S} = \{0,1\}$. Thus, the observational unit for the target sample is $O = (Y, A, \boldsymbol{X}, S)$.

The data generating processes which result in randomized and observational data realizations differ. The randomized data consist of $n_{\text{RCT}}$ i.i.d. realizations conditional on selection into the randomized group, $S = 1$. The observational unit for the randomized data is $O_{\text{RCT}} = (Y, A, \boldsymbol{X}, S = 1) \sim (Y, A, \boldsymbol{X} | S = 1) \equiv P_{\text{RCT}}$. Similarly, the observational data consist of $n_{\text{obs}}$ i.i.d. draws conditional on selection into the observational study, $S = 0$. The observational unit for the observational data is thus $O_{\text{obs}} = (Y, A, \boldsymbol{X}, S = 0) \sim (Y, A, \boldsymbol{X} | S = 0) \equiv P_{\text{obs}}$.

### 2.2.2 Estimand

As per the potential outcomes framework, let $Y^a$ be the potential outcome if intervention $a$ were assigned. The estimands of interest for our intervention are the target population

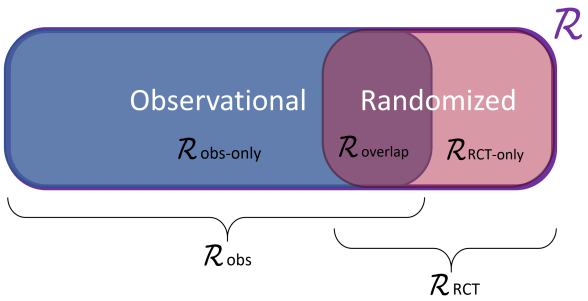



**Figure (2.1)** *Overlap and nonoverlap regions in the target population*

*The region of support in the target population's covariate distribution, $\mathcal{R}$, is the union of the randomized group's support ($\mathcal{R}_{RCT}$) and the observational group's support ($\mathcal{R}_{obs}$). Partial overlap exists between $\mathcal{R}_{RCT}$ and $\mathcal{R}_{obs}$: $\mathcal{R}_{overlap}$ corresponds to the region of overlap (i.e., region of common support in the covariate distributions) between data sources; $\mathcal{R}_{obs\text{-}only}$ corresponds to the region only represented in the observational data and $\mathcal{R}_{RCT\text{-}only}$ corresponds to the region only represented in the randomized data.*

treatment-specific means (PTSMs): $E(Y^a)$ for $\forall a \in \mathcal{A}$, as have been explored in prior analyses with multiple unordered treatments (Rose and Normand, 2019). In contrast, study treatment-specific means (STSMs) are mean counterfactual outcomes for a given treatment over a given study population: $E(Y^a|S=s)$ for $\forall a \in \mathcal{A}, s \in \mathcal{S}$. Because no given health plan serves as a natural "control" comparator, treatment-specific means rather than the target population average treatment effect (PATE: $E(Y^a) - E(Y^{a'})$) are of interest.

### 2.2.3 Defining and determining overlap and nonoverlap regions

Covariate distributions differ between randomized and observational groups: $P(\boldsymbol{X}|S=1) \neq P(\boldsymbol{X}|S=0)$. Furthermore $P(\boldsymbol{X}=\mathbf{x}|S=0)=0$ and $P(\boldsymbol{X}=\mathbf{x}'|S=1)=0$ for some $\mathbf{x},\mathbf{x}' \in \mathcal{R}$. Namely, a portion of the observational data is not well-represented in the randomized data and potentially a portion of the randomized data may not be well-represented in the observational data. However, there is a region of overlap between randomized and observational covariate distributions. Overlap refers to common support across randomized and observational populations in the distribution of outcome predictors associated with study selection (or effect modifiers associated with study selection if the estimand of interest had been an average treatment effect): $\mathcal{R}_{\text{overlap}} = \mathbf{x} \in \mathcal{R} : P(\boldsymbol{X}=\mathbf{x}|S=1) > 0 \cap P(\boldsymbol{X}=\mathbf{x}|S=0) > 0$ (Figure 2.1). Regions of nonoverlap therefore correspond to regions of the covariate distribution where

either only observational individuals ($\mathcal{R}_{\text{obs-only}}$) or only randomized individuals ($\mathcal{R}_{\text{RCT-only}}$) would be observed, i.e., regions where units in one study population are not eligible to be in the other study population.

The target sample covariate distribution ($\mathcal{R}$) can therefore be decomposed as: $\mathcal{R} = \mathcal{R}_{\text{overlap}} \cup \mathcal{R}_{\text{obs-only}} \cup \mathcal{R}_{\text{RCT-only}}$ Thus, $\mathcal{R}_{\text{obs}} = \mathcal{R}_{\text{overlap}} \cup \mathcal{R}_{\text{obs-only}}$ and $\mathcal{R}_{\text{RCT}} = \mathcal{R}_{\text{overlap}} \cup \mathcal{R}_{\text{RCT-only}}$. $\mathcal{R}_{\text{RCT-only}}$ and $\mathcal{R}_{\text{obs-only}}$ may be null sets. Let $R$ be an indicator for being in the respective region, e.g., $R_{\text{overlap}} = \mathbb{1}(\text{membership in } \mathcal{R}_{\text{overlap}})$.

At times, it may be the case that rather than a union of a randomized and observational study being representative of the target population, a reweighted union of the two studies may be representative, such as when working with a random sample of observational data for computational efficiency (which we do for our analysis), or when data is collected through survey sampling. In this case, one can, through reweighting, map the randomized and observational study regions of covariate support, $\mathcal{R}_{\text{RCT}}$ and $\mathcal{R}_{\text{obs}}$, into a transformation, $\mathcal{R}_{\text{RCT}} \to \mathcal{R}^*_{\text{RCT}}$ and $\mathcal{R}_{\text{obs}} \to \mathcal{R}^*_{\text{obs}}$, in which the decomposition above of $\mathcal{R}^* = \mathcal{R}^*_{\text{overlap}} \cup \mathcal{R}^*_{\text{obs-only}} \cup \mathcal{R}^*_{\text{RCT-only}}$ holds. Note that this includes the possibility of the target population being represented by just the observational data.

While the above definition of overlap corresponds to a population feature, nonoverlap can also occur due to having a finite sample; by chance, the data may be sparse in some region of the covariate distribution even though that region has support. In practice, we will account for overlap as both a population and sample feature, determining regions of the covariate space that have common support and observed data from both groups. To estimate the region of overlap, $\mathcal{R}_{\text{overlap}}$, we extend a data-driven approach for determining areas of treatment overlap based on propensity scores for treatment assignment (Nethery *et al.*, 2018). We adopt a similar approach for the propensity score for study selection $\pi_S = P(S|\boldsymbol{X})$, but on the logit scale to give more granularity to very low and very high propensity scores.

## 2.3 Assumptions and identification

### 2.3.1 Standard assumptions

Identifying population causal quantities such as PTSMs and PATEs standardly relies on the following sufficient generalizability assumptions (Stuart *et al.*, 2011; Tipton, 2013a; Degtiar and Rose, 2021):

**Internal validity**

1. Conditional treatment exchangeability: $Y^a \perp A \,|\boldsymbol{X}, S = s$ for all $a \in \mathcal{A}, s \in \mathcal{S}$
2. Positivity of treatment assignment: $P(\boldsymbol{X} = \boldsymbol{x}|S = 1) > 0 \;\Rightarrow P(A = a|\boldsymbol{X} = \boldsymbol{x}, S = 1) > 0$ with probability 1 for all $a \in \mathcal{A}$.
3. Stable unit treatment value assumption (SUTVA) for treatment assignment: if $A_i = a$ then $Y_i = Y_i^a$

**External validity**

4. Conditional exchangeability for study selection: $Y^a \perp S \,|\boldsymbol{X}$ for all $a \in \mathcal{A}$
5. Positivity of study selection: $P(\boldsymbol{X} = \boldsymbol{x}) > 0 \;\Rightarrow P(S = s|\boldsymbol{X} = \boldsymbol{x}) > 0$ with probability 1 for all $s \in \mathcal{S}$
6. SUTVA for study selection: if $S_i = s$ and $A_i = a$ then $Y_i = Y_i^a$

For a specific estimand of interest, Assumptions 1 and 4 can be weakened. For example, when estimating PTSMs, Assumptions 1 and 4 can be replaced by the following:

1. Mean conditional treatment exchangeability: $E(Y^a|A = a, S = s, \boldsymbol{X}) = E(Y^a|S = s, \boldsymbol{X})$ for all $a \in \mathcal{A}, s \in \mathcal{S}$
4. Mean conditional exchangeability for study selection: $E(Y^a|S = s, \boldsymbol{X}) = E(Y^a|\boldsymbol{X})$ for all $a \in \mathcal{A}, s \in \mathcal{S}$

### 2.3.2 Relaxation of the mean conditional treatment exchangeability and positivity of study selection assumptions

To accommodate the potential violations of standard assumptions 1 and 5 for PTSMs, we replace these assumptions with the following relaxations:

1b. Mean conditional exchangeability in the randomized group:

$$E(Y^a|S=1,A=a,\boldsymbol{X}) = E(Y^a|S=1,\boldsymbol{X}) \text{ for all } a \in \mathcal{A}$$

and constant conditional bias in the observational group:

$$E(Y^a|S=0,A=a,\boldsymbol{X}) - E(Y^a|S=1,A=a,\boldsymbol{X})$$
$$= E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})$$

5b. Overlap between study samples: there exists a non-null set $\mathcal{R}_{\text{overlap}}$ such that $P(\boldsymbol{X}=\boldsymbol{x}|R_{\text{overlap}}) > 0 \Rightarrow P(S=s|\boldsymbol{X}=\boldsymbol{x}) > 0$ with probability 1 for all $s \in \mathcal{S}$.

Assumption 1b corresponds to the same conditional bias relationship holding in $\mathcal{R}_{\text{overlap}}$ as $\mathcal{R}_{\text{obs}}$: $b(a,\boldsymbol{x}) = b(a,\boldsymbol{x}|R_{\text{overlap}}=1)$, for all $a \in \mathcal{A}$ where $b(a,\boldsymbol{x}) \equiv E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=0,\boldsymbol{X}=\boldsymbol{x})$. See Appendix B.1 in the supplementary material for a derivation and further motivation for these weakened identifiability assumptions, in addition to a restatement of Assumption 1b with respect to the unmeasured confounders that are implicitly being integrated over.

More specifically (and more weakly), Assumption 1b must hold in expectation over the $\boldsymbol{X}$ covariate distribution in the observational data (mean constant conditional bias):

$$E_{\boldsymbol{X}}\Big\{E(Y^a|S=0,A=a,\boldsymbol{X}) - E(Y^a|S=1,A=a,\boldsymbol{X})\Big|S=0\Big\}$$
$$= E_{\boldsymbol{X}}\Big\{E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big|S=0\Big\}$$

Assumption 1b states that the relationship between bias and measured covariates is unrelated to being in the overlap vs. nonoverlap regions, i.e., that the distribution of unmeasured confounders does not differ between $\mathcal{R}_{\text{overlap}}$ and $\mathcal{R}_{\text{obs}}$, conditioning on $\boldsymbol{X}$ and

$A$. This assumption is strictly weaker than the no unmeasured confounding assumption in that the assumption of no unmeasured confounding is nested within Assumption 1b: with no unmeasured confounding, $b(a, \boldsymbol{x}) = 0$. Constant conditional bias can be seen as an extension of the assumption made for cross-design synthesis (Kaizar, 2011), except that constant conditional bias is allowed to depend on measured covariates and a dichotimization of the covariate distribution support into overlap and nonoverlap regions replaces predefined eligibility determining overlap region membership. We hence assume that the covariates $\boldsymbol{X}$ capture all factors that would lead to differential bias in the overlap as nonoverlap regions. This suggests that we can estimate bias in the overlap region and use those estimates to also extrapolate to and correct for bias in the observational group's nonoverlap region.

Assumption 1b is untestable, just as is the assumption of no unmeasured confounding; it would fail if the processes that drove unmeasured confounding differed between overlap and nonoverlap regions in a way that was not captured by measured covariates, or if the distribution of the unmeasured confounder differed between those regions in such a way as to create different conditional expectation relationships. This could occur, for example, if an unmeasured confounder drove overlap region membership. If the constant bias assumption is not reasonable for a given setting, one can alternatively perform sensitivity analysis to obtain bounds on PTSMs (Appendix B.2).

In practice, Assumption 5b's region of overlap should be sufficiently large to learn the bias term, i.e., sufficiently large for Assumption 1b to hold. Empirical violations of Assumption 5b are partially testable using $\pi_S$; the existence of overlap in the propensity score distributions between randomized and observational groups provides evidence for this assumption. Observational group propensity scores may also be close to zero and lacking overlap with randomized group propensity scores when the observational group size far exceeds the randomized group.

Of note, the $\boldsymbol{X}$ needed for Assumptions 1b and 2 and the $\boldsymbol{X}$ needed for Assumption 4 and 5b may differ. As a result, the region of overlap should exist with respect to outcome predictors, but should be large enough to ensure that Assumption 1b holds. It is therefore

reasonable to use an $\boldsymbol{X}$ matrix that contains all outcome predictors and confounders to assess all assumptions. Thus, as described earlier, our $\boldsymbol{X}$ is the union of the covariates sets needed for all assumptions to hold.

### 2.3.3 Identification

Under the modified assumptions above, the causal estimand of interest can be identified by the following CCDS functional of the observed data:

$$\begin{aligned}\psi_{\text{CCDS}}(a) =& E_{\boldsymbol{X}|S=1}\Big[E(Y|S=1,A=a,\boldsymbol{X})\Big|S=1\Big]P(S=1)\\ &+E_{\boldsymbol{X}|S=0}\Big[E(Y|S=0,A=a,\boldsymbol{X})\Big|S=0\Big]P(S=0)\\ &-E_{\boldsymbol{X}|S=0}\Big[\Big\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\\ &\quad -E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\}\Big|S=0\Big]P(S=0)\end{aligned}$$

See Appendix B.3 for the proof and Appendix B.4 for alternative functionals that identify the PTSM, derived through different decompositions of the data.

## 2.4 Estimators

We develop four novel estimators that combine randomized and observational data to estimate PTSMs relying on our CCDS framework. The novel estimators consist of outcome regression, 2-stage outcome regression, inverse probability weighting, and double robust augmented inverse probability weighting approaches.

### 2.4.1 CCDS outcome regression estimator

The CCDS outcome regression (CCDS-OR) estimator uses outcome regressions to estimate the combination of the conditional distributions in $\psi_{\text{CCDS}}(a)$:

$$\begin{aligned}\hat{\psi}_{\text{CCDS-OR}}(a) =& \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(S_i=1, A_i=a, \boldsymbol{X}_i)\mathbb{1}(S_i=1) + \hat{Q}(S_i=0, A_i=a, \boldsymbol{X}_i)\mathbb{1}(S_i=0) \\ & - \Big\{\hat{Q}(S_i=0, A_i=a, \hat{R}_{\text{overlap, i}}=1, \boldsymbol{X}_i) \\ & \quad - \hat{Q}(S_i=1, A_i=a, \hat{R}_{\text{overlap, i}}=1, \boldsymbol{X}_i)\Big\}\mathbb{1}(S_i=0).\end{aligned}$$

where $\hat{R}_{\text{overlap}}$ is estimated as described in Section 2.2.3, $\hat{Q}(S=1, A=a, \boldsymbol{X})$ is an estimator for $E(Y|S=1, A=a, \boldsymbol{X})$, $\hat{Q}(S=0, A=a, \boldsymbol{X})$ is an estimator for $E(Y|S=0, A=a, \boldsymbol{X})$, $\hat{Q}(S=0, A=a, \hat{R}_{\text{overlap}}=1, \boldsymbol{X})$ is an estimator of $E(Y|S=0, A=a, R_{\text{overlap}}=1, \boldsymbol{X})$, and $\hat{Q}(S=1, A=a, \hat{R}_{\text{overlap}}=1, \boldsymbol{X})$ is an estimator of $E(Y|S=1, A=a, R_{\text{overlap}}=1, \boldsymbol{X})$. The first term corresponds to treatment specific mean estimates for the randomized subset of the target sample, the second term provides preliminary estimates for the observational subset of the target sample, and the third term debiases the preliminary observational data estimates. Implementation considerations for regression choices and a conditional treatment-specific mean version of the estimator are presented in Appendix B.5.

### 2.4.2 2-stage CCDS outcome regression estimator

To avoid overfitting to overlap region trends, the 2-stage CCDS estimator replaces the debiasing term, the third term, with a 2-stage regression:

$$\begin{aligned}\hat{\psi}_{\text{2-stage CCDS}}(a) =& \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(S_i=1, A_i=a, \boldsymbol{X}_i)\mathbb{1}(S_i=1) + \hat{Q}(S_i=0, A_i=a, \boldsymbol{X}_i)\mathbb{1}(S_i=0) \\ & - \hat{b}(S_i=1, a, \boldsymbol{X}_i)\mathbb{1}(S_i=0)\end{aligned}$$

where $\hat{b}(S_i=1, a, \boldsymbol{X}_i)$ is estimated via

(1) $\hat{b}'(S_i=1, a, \boldsymbol{X}_i) = \hat{Q}(S_i=0, A_i=a, \hat{R}_{\text{overlap, i}}=1, \boldsymbol{X}_i)\mathbb{1}(S_i=1, \hat{R}_{\text{overlap, i}}=1) - \hat{Q}(S_i=$

$1, A_i = a, \hat{R}_{\text{overlap, i}} = 1, \boldsymbol{X}_i) \mathbb{1}(S_i = 1, \hat{R}_{\text{overlap, i}} = 1)$

(2) $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i) = \frac{\hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i)} \hat{g}(\boldsymbol{X}_i)$ with

$$\hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i) = \frac{\mathbb{1}(S_i = 1, \hat{R}_{\text{overlap, i}} = 1) \hat{P}(S_i = 0 | \boldsymbol{X}_i)}{\hat{P}(\hat{R}_{\text{overlap, i}} = 1 | S_i = 1, \boldsymbol{X}_i) \hat{P}(S_i = 1 | \boldsymbol{X}_i)}$$

and $\hat{g}(\boldsymbol{X})$ an estimator of a regression function described below.

Namely, Stage (1), estimates an intermediate bias term $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i)$ using randomized overlap data: bias estimates are the difference in predicted counterfactual outcomes using regressions fit to the overlap region of the observational vs. randomized data, creating predictions for the randomized overlap data. As there is no bias in expectation in the randomized overlap data, any estimated bias stems from the regression $\hat{Q}(S_i = 0, A_i = a, \boldsymbol{X}_i)$. Stage (2) then fits a weighted regression with the estimates of $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i)$ from Stage (1) as the outcome. This second stage focuses on the relationship between the bias estimates in the overlap region and measured covariates. The debiasing term $\hat{b}(S_i = 1, a, \boldsymbol{X}_i)$ is then estimated for the observational data from the fixed regression fit $\hat{g}(\boldsymbol{X})$ in Stage (2).

The weight, $\hat{w}_{\text{bias}}$, standardizes the randomized data to the observational data so that the bias term is estimated for the covariate distribution of interest. The weights follow from $P(S = 0) = E\Big[P(S = 0 | \boldsymbol{X})\Big] = E\Big[\mathbb{1}(S = 1, R_{\text{overlap}} = 1) P(S = 0 | \boldsymbol{X}) / \Big(P(R_{\text{overlap}} = 1 | S = 1, \boldsymbol{X}) P(S = 1 | \boldsymbol{X})\Big)\Big]$. Reweighting will frequently not face issues when positivity of selection violations occur because $S = 1$ data is used to estimate the bias term and thus should not have many values close to zero for $\hat{P}(S_i = 1 | \boldsymbol{X}_i)$, which is in the denominator of the weight. Thus, while weighting is not required in such a 2-stage approach, the weights add robustness compared to an unweighted approach without common drawbacks of weighting, such as variance inflation due to unstable weights.

Appendix B.6 presents a 2-stage approach that does not restrict itself to the overlap region (2-stage whole data), which suffers from the same reliance on extrapolating beyond randomized group support as does using only the randomized data, highlighting the importance of focusing on the overlap region to debias observational data.

### 2.4.3 CCDS inverse probability weighting estimator

The cross-design synthesis inverse probability weighting (CCDS-IPW) estimator with stabilized weights uses propensity models to estimate PTSMs (see Appendix B.7 for the proof):

$$\begin{aligned}\hat{\psi}_{\text{CCDS-IPW}}(a) = \frac{n_{\text{rand}}}{n}\left[\sum_{i=1}^{n}\hat{w}_1(S_i,A_i,\boldsymbol{X}_i)\right]^{-1}\sum_{i=1}^{n}\hat{w}_1(S_i,A_i,\boldsymbol{X}_i)Y_i \\ + \frac{n_{\text{obs}}}{n}\left[\sum_{i=1}^{n}\hat{w}_2(S_i,A_i,\boldsymbol{X}_i)\right]^{-1}\sum_{i=1}^{n}\hat{w}_2(S_i,A_i,\boldsymbol{X}_i)Y_i \\ - \frac{n_{\text{obs}}}{n}\left\{\left[\sum_{i=1}^{n}\hat{w}_3(S_i,A_i,\boldsymbol{X}_i)\right]^{-1}\sum_{i=1}^{n}\hat{w}_3(S_i,A_i,\boldsymbol{X}_i)Y_i \right. \\ \left. - \left[\sum_{i=1}^{n}\hat{w}_4(S_i,A_i,\boldsymbol{X}_i)\right]^{-1}\sum_{i=1}^{n}\hat{w}_4(S_i,A_i,\boldsymbol{X}_i)Y_i\right\}\end{aligned}$$

where:

$$\begin{aligned}\hat{w}_1(S_i,A_i,\boldsymbol{X}_i) &= \frac{\mathbb{1}(S_i=1,A_i=a)}{\hat{P}(A_i=a|S_i=1,\boldsymbol{X}_i)} \\ \hat{w}_2(S_i,A_i,\boldsymbol{X}_i) &= \frac{\mathbb{1}(S_i=0,A_i=a)}{\hat{P}(A_i=a|S_i=0,\boldsymbol{X}_i)} \\ \hat{w}_3(S_i,A_i,\boldsymbol{X}_i) &= \frac{\mathbb{1}(S_i=0,A_i=a,\hat{R}_{\text{overlap, i}}=1)}{\hat{P}(\hat{R}_{\text{overlap, i}}=1|S_i=0,\boldsymbol{X}_i)\hat{P}(A_i=a|S_i=0,\hat{R}_{\text{overlap, i}}=1,\boldsymbol{X}_i)} \\ \hat{w}_4(S_i,A_i,\boldsymbol{X}_i) &= \frac{\mathbb{1}(S_i=1,A_i=a,\hat{R}_{\text{overlap, i}}=1)[1-\hat{P}(S_i=1|\boldsymbol{X}_i)]}{\hat{P}(S_i=1|\boldsymbol{X}_i)\hat{P}(\hat{R}_{\text{overlap, i}}=1|S_i=1,\boldsymbol{X}_i)\hat{P}(A_i=a|S_i=1,\hat{R}_{\text{overlap, i}}=1,\boldsymbol{X}_i)}\end{aligned}$$

Here, positivity of selection violations will often not lead to unstable weights since $\hat{P}(\hat{R}_{\text{overlap, i}} = 1|S_i = 1,\boldsymbol{X}_i)\hat{P}(S_i = 1|\boldsymbol{X}_i)$ only appears in the denominator for $\hat{w}_4$; these individuals, by overlap region construction, have propensity scores for selection bounded away from zero. Normalizing weights by their sum adds stability (Robins *et al.*, 2000). Nonetheless, this method may face lack of efficiency and potentially unstable estimates, particularly from estimating the second bias term contribution weighted by $\hat{w}_4$, as the components are estimated using small subsets of the data relative to the overall sample—only individuals randomized in the overlap region on a given treatment arm. This problem is exacerbated with many treatment groups, particularly for rare treatments.

### 2.4.4 CCDS augmented inverse probability weighting estimator

Our double robust estimator provides consistent estimates when either the outcome regressions or product of propensity regressions are correctly specified in each of the terms of $\psi_{\text{CCDS}}(a)$. The CCDS augmented inverse probability weighted (CCDS-AIPW) estimator is as follows:

$$
\begin{aligned}
&\hat{\psi}_{\text{CCDS-AIPW}}(a) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{n_{\text{rand}}}{n} \frac{\hat{w}_1(S_i, A_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_1(S_i, A_i, \boldsymbol{X}_i)} \Big\{ Y_i - \hat{Q}_i(S=1, A=a, \boldsymbol{X}) \Big\} + \mathbb{1}(S_i = 1)\hat{Q}_i(S=1, A=a, \boldsymbol{X}) \\
&\quad + \frac{n_{\text{obs}}}{n} \frac{\hat{w}_2(S_i, A_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_2(S_i, A_i, \boldsymbol{X}_i)} \Big\{ Y_i - \hat{Q}_i(S=0, A=a, \boldsymbol{X}) \Big\} + \mathbb{1}(S_i = 0)\hat{Q}_i(S=0, A=a, \boldsymbol{X}) \\
&\quad - \frac{n_{\text{obs}}}{n} \frac{\hat{w}_3(S_i, A_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_3(S_i, A_i, \boldsymbol{X}_i)} \Big\{ Y_i - \hat{Q}_i(S=0, A=a, \hat{R}_{\text{overlap}} = 1, \boldsymbol{X}) \Big\} \\
&\qquad - \mathbb{1}(S_i = 0)\hat{Q}_i(S=0, A=a, \hat{R}_{\text{overlap}} = 1, \boldsymbol{X}) \\
&\quad + \frac{n_{\text{obs}}}{n} \frac{\hat{w}_4(S_i, A_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_4(S_i, A_i, \boldsymbol{X}_i)} \Big\{ Y_i - \hat{Q}_i(S=1, A=a, \hat{R}_{\text{overlap}} = 1, \boldsymbol{X}) \Big\} \\
&\qquad + \mathbb{1}(S_i = 0)\hat{Q}_i(S=1, A=a, \hat{R}_{\text{overlap}} = 1, \boldsymbol{X})
\end{aligned}
$$

with $\hat{w}(S_i, A_i, \boldsymbol{X}_i)$ and $\hat{Q}_i(S, A, \boldsymbol{X})$ as defined above. CCDS-AIPW is a double robust estimator that is asymptotically efficient when the propensity and outcome regressions are estimated consistently. See Appendix B.8 for a derivation of the efficient influence function.

### 2.4.5 Inference

Confidence intervals and standard errors in our machine-learning-based analyses were calculated using a nonparametric bootstrap (Efron and Tibshirani, 1994). When using parametric regressions, a sandwich variance approach can be used to derive sampling variance, following M-estimation theory.

### 2.4.6 Comparison estimators

There are no existing methods that address both the overlap and unmeasured confounding challenges specific to our data setting. While the estimator of Rosenman *et al.* (2020) addresses overlap and unmeasured confounding, it assumes that treatment effects are identical between randomized and observational groups within the same stratum of effect modifiers, which is unlikely to hold in our setting. We therefore compare against two simple approaches. The first (rand estimator) fits an outcome regression using randomized data to extrapolate to the entire target population—including outside its region of support (Kern *et al.*, 2016): $\hat{\psi}_{\text{rand}}(a) = 1/n \sum_{i=1}^{n} \hat{Q}(S_i = 1, A_i = a, \boldsymbol{X}_i)$. This extrapolation may yield bias when the relationship between covariates and potential outcomes differs in $\mathcal{R}_{\text{overlap}}$ compared to $\mathcal{R}_{\text{obs}}$ in a way that cannot be extrapolated from the randomized data. This second (obs/rand estimator) is similar in spirit to Kern *et al.* (2016) and Prentice *et al.* (2006), though those fit one outcome regression to both randomized and observational data and estimate effects for just the observational data or just the randomized data, respectively. The obs/rand estimator we deploy here fits an outcome regression using randomized data to estimate counterfactuals for the randomized data and fits an outcome regression using observational data to estimate counterfactuals for the observational data: $\hat{\psi}_{\text{obs/rand}}(a) = 1/n \sum_{i=1}^{n} \hat{Q}_i(S = 1, A = a, \boldsymbol{X}) \mathbb{1}(S_i = 1) + \hat{Q}_i(S = 0, A = a, \boldsymbol{X}) \mathbb{1}(S_i = 0)$. This approach assumes there is no unmeasured confounding in the observational data. We used outcome regressions for rand and obs/rand estimators rather than approaches that incorporate propensities for selection as the latter will result in denominators close to zero due to lack of overlap.

## 2.5 Simulation studies

We designed a broad series of simulations to evaluate the finite sample performance of our novel CCDS estimators compared to alternative approaches for estimating PTSMs as well as the PATE, examining two treatment groups $A \in \{1, 2\}$. We assessed performance in the

presence of (1) complex data-generating mechanisms such that the randomized data do not extrapolate well outside their support, (2) unmeasured confounding in the observational data, and (3) positivity of selection violations. We also studied alternative data generating processes including different sample sizes, constant bias violations, unmeasured confounding settings, overlap settings, ratios of $n_{\text{RCT}}$ to $n_{\text{obs}}$, positivity of selection violation settings, exchangeability of study selection violations, overlap region determination settings, propensity for selection relationships, alternative outcome models, and alternative regression fits. In total, we examined 84 different data generating scenario × regression choice combinations.

In the base case, we generated a target population of 1 million individuals from which we drew random samples of size $n = 10,000$, with data generating mechanism $P(Y,S,A,\boldsymbol{X},U) = P(\boldsymbol{X})P(U|\boldsymbol{X})P(S|\boldsymbol{X},U)P(A|S,\boldsymbol{X},U)P(Y|S,A,\boldsymbol{X},U)$. The data had four independent measured confounders $X_1,...,X_4 \sim N(0,1)$; an unmeasured confounder $U \sim Binom(0.5)$; selection into the randomized group driven by the strongest confounder such that there existed $\mathcal{R}_{\text{RCT-only}}$ ($S = 1$ if $X_1 > QNnorm(0.9)$), $\mathcal{R}_{\text{obs-only}}$ ($S = 0$ if $X_1 < QNorm(0.5)$), and $\mathcal{R}_{\text{overlap}}$ ($S \sim Binom(0.5)$, otherwise). This study selection process resulted in approximately a 1:4 ratio of randomized to observational individuals. Treatment assignment was $A \sim Binom(0.6)$ for $S = 1$ and $A \sim Binom(\text{logit}^{-1}(-0.8+0.125X_1+0.1X_2+0.075X_3+0.05X_4+0.1(X_1+1)^3+0.625U))$ for $S = 0$. The outcome was generated from the same distribution for both groups, $Y \sim Norm(\mu_Y,1)$, where $\mu_Y = -1.5-3A+4X_1+4X_2+3X_3+2X_4+0.4(X_1+1)^3+4AX_1+10U$.

Estimators were fit with linear outcome regressions as well as an ensemble of 8 machine learning approaches. We implemented 2000 simulation iterations and 1000 bootstrap replications to generate confidence intervals. Propensities and their products used in weight denominators were trimmed at 0.001. We implemented the simulations in R, including the `SuperLearner` package (Polley *et al.*, 2019) and the `pw_overlap` function for overlap region estimation (Nethery *et al.*, 2018). See Appendix B.9 for the correspondence of our simulation design with identifiability assumptions, descriptions of alternative data generating mechanisms, and further implementation details. Our code is available on GitHub (https://github.com/idegtiar1).

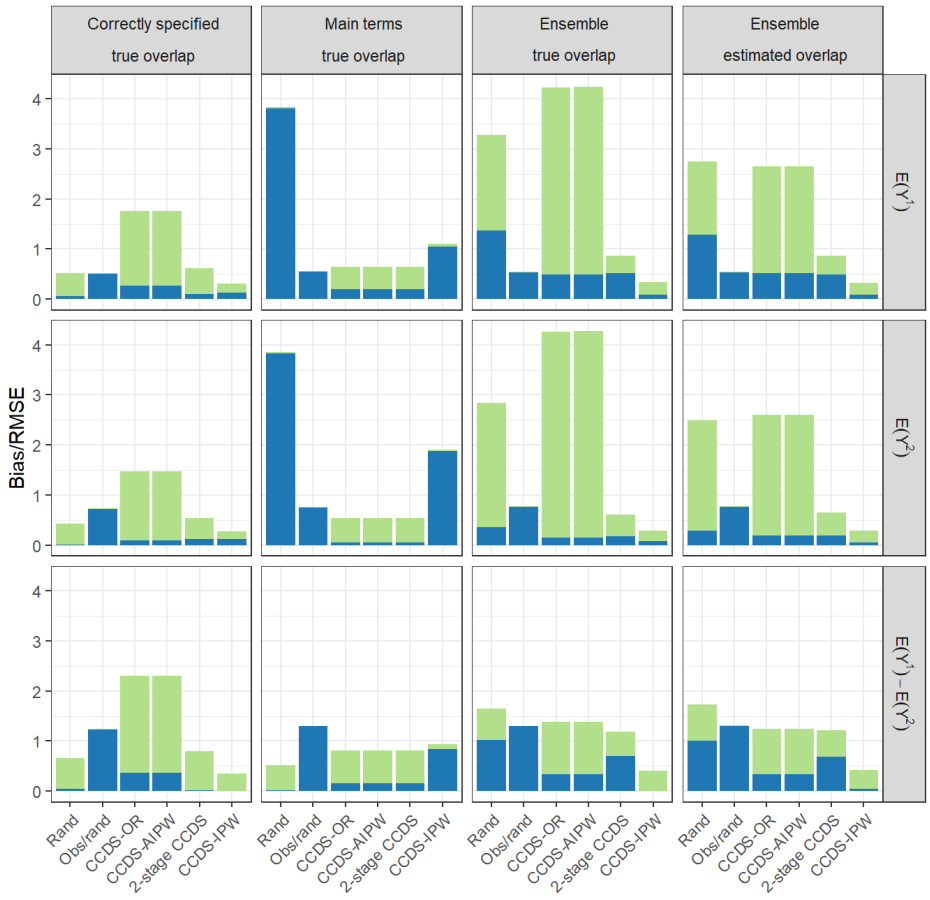



**Figure (2.2)** *Bias and RMSE for PTSM and PATE estimates for* $n = 10,000$ *across 2000 simulation iterations and 1000 bootstrap replications*

*Absolute bias is the darker portion of each bar; RMSE corresponds to the total bar size.*

**Main Findings.** Results across different regression specifications highlight the estimators' relative strengths and disadvantages (Figure 2.2). At the base case sample size of $n = 10,000$, CCDS-OR and CCDS-AIPW performance was almost identical. These estimators suffered from large variability when fitting complex regressions in a small overlap region, which we observed in the correctly specified and ensemble settings. In contrast, the CCDS-OR and CCDS-AIPW estimators showed little bias and variance when fitting underspecified main terms regressions; underspecification avoids overfitting in a small overlap region. The 2-stage CCDS estimator decreased bias and variance when using correctly specified or ensemble

regressions, relative to the (1-stage) CCDS-OR estimator and CCDS-AIPW. In the main terms setting, its estimates were identical to those of CCDS-OR due to linearity and additivity.

The CCDS-IPW had the smallest bias and RMSE throughout all settings except when fitting main terms regressions where it grossly misspecifies the propensity for selection, resulting in large remnant bias for that setting. However, the estimator's efficiency was due to the outcome model having more variability compared to the propensity models; e.g., the propensity for selection was deterministically assigned by $X_1$. With a more probabilistic relationship and smaller propensity scores, CCDS-IPW's bias and variance increased.

While the rand estimator performed well with correctly specified regressions, using only main terms regressions resulted in large bias due to poor extrapolation beyond its support. With more flexible ensemble approaches, the rand estimator suffered from both large bias and large variance. The obs/rand estimator was subject to unmeasured confounding bias, which existed even when correctly specified regressions were fit, though it had relatively low RMSE due to the large observational sample size.

**Estimating Overlap.** The last column of Figure 2.2 presents results from overlap region estimation using $\alpha = 0.01 \times \text{range}(\text{logit}(\pi_S))$ and $\beta = 0.01 \times \min(n_{\text{RCT}}, n_{\text{obs}})$. The region of overlap consists of points in the logit of the propensity score for selection that have at least $\beta$ observations from each study group within an interval of size $\alpha$ around that point (Nethery *et al.*, 2018). With these specifications, compared to the truth, the estimated overlap region had a similar number of observational individuals (38% vs. 35%) and randomized individuals (50% vs. 48%). Performance was similar or better when estimating the overlap region in this setting and across the various other data generating mechanisms and overlap region hyperparameter specifications we examined (Appendix B.9).

**Coverage.** The obs/rand estimator showed 0% coverage across all settings while all CCDS estimators were able to achieve nominal coverage, except the CCDS-IPW estimator when using grossly misspecified linear regressions (Figure 2.3). The rand estimator attained 0% coverage in the main terms setting for the PTSMs but 95% coverage for the PATE due to linear regressions correctly specifying treatment effects but not treatment-specific means in

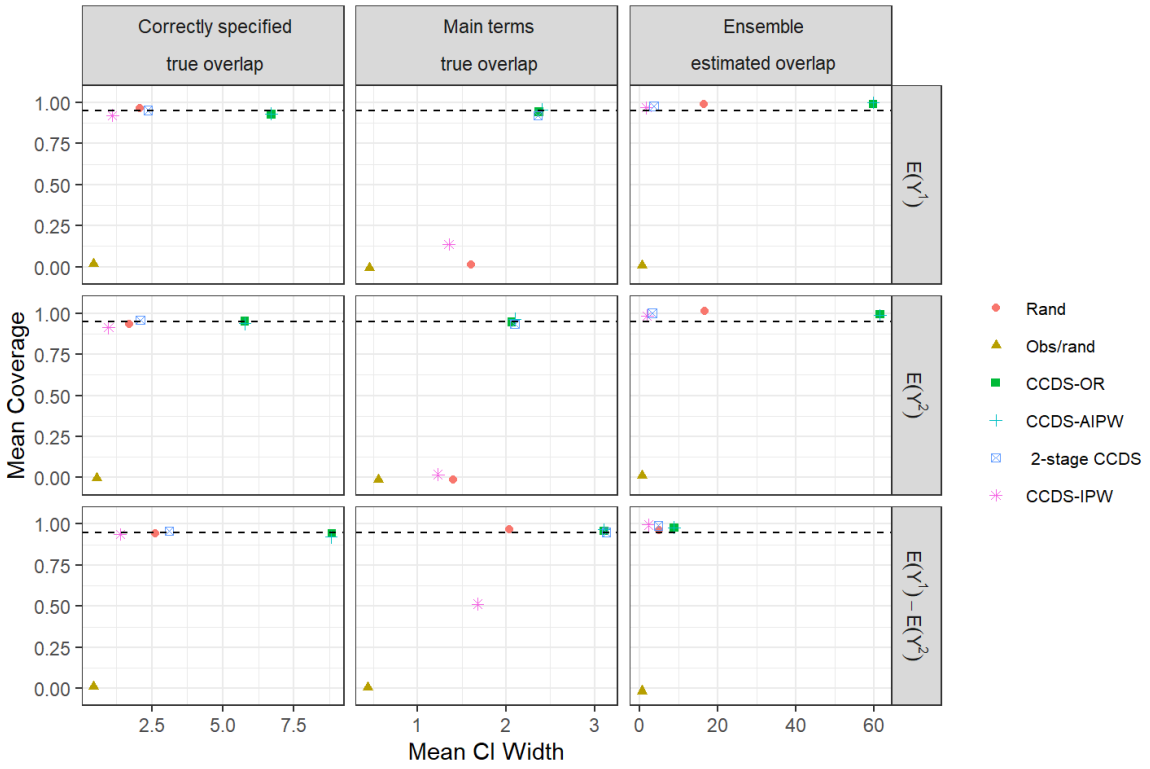



**Figure (2.3)** *Coverage and confidence interval width for PTSM and PATE estimates for* $n = 10,000$ *across 2000 simulation iterations and 1000 bootstrap replications*

*The dashed line corresponds to the target coverage of 95%.*

this data-generating mechanism; coverage remains low when the PATE does not extrapolate well from the randomized data. Thus, while the bias and RMSE of the CCDS estimators may or may not decrease compared to the obs/rand estimator (as shown in Figure 2.2) due to remnant estimation error from misspecifying regressions, which is particular evident with ensemble approaches, the poor coverage of the obs/rand estimator indicates this can be a false indication of precision.

**Alternative Data Generating Mechanisms.** CCDS estimator bias and RMSE shrunk with more overlap and with increasing proportions of randomized data. As unmeasured confounding bias increased, there was no corresponding increase in bias across CCDS estimators with correctly specified regressions and only a slight increase with ensembles. However, variance increased, reflecting additional uncertainty in settings with more unmeasured confounding.

Violating the constant conditional bias assumption increased bias for the rand and all CCDS estimators, with CCDS estimators generally performing better than the rand estimator. All estimators performed poorly when the exchangeability of study selection assumption was violated. The RMSE for CCDS-IPW was most impacted by a smaller ratio of randomized to observational individuals. Overall, results for the CCDS estimators were similar across alternative data generating mechanisms. Further details can be found in Appendix B.9.

## 2.6 Medicaid study

Medicaid, administered by the Centers for Medicare & Medicaid Services, provides insurance for low-income and disadvantaged Americans, covering a fifth of all individuals in the United States (Centers for Medicare & Medicaid Services, 2020). As described earlier, MMC provides health insurance plans for all but certain exempt groups (Medicaid, 2020), and beneficiaries who do not actively choose a health plan are randomized to one. Understanding the impact of these individual MMC health plans on health care spending is an open question. However, generalizing the 7% of beneficiaries who are randomized to the full NYC Medicaid population may be hampered by a lack of overlap in parts of the covariate distributions between randomized and observational (active chooser) groups. Yet, data from observational beneficiaries may be subject to potential unmeasured confounding from variables not captured in the claims data.

We estimated the causal effects of enrollment into NYC MMC health plans on health care spending for all NYC Medicaid beneficiaries with at least 6 months of follow-up, applying our novel CCDS and comparison estimators. Health care spending was examined over 6 months on the log scale, as log(spending + 1), adjusting for baseline spending decile, age, documented sex, aid group, whether the beneficiary received social security income, neighborhood, and neighborhood poverty level. Further descriptions of the data can be found in Geruso *et al.* (2020). We used all 65,591 randomized beneficiaries and a 10% random subset of observational beneficiaries within the study period (2008 - 2012) for computational efficiency, which totaled 98,232. For the 1% of beneficiaries with missing

baseline spending, baseline spending was imputed to be zero (the most likely reason for missingness was no spending) along with an indicator for missingness. Regressions were fit using a `SuperLearner` ensemble (of `glm`, `glmnet` with $\alpha = 0.5$, `gam`, and `nnet`). Propensity scores and their products used in weight denominators were trimmed at 0.001. To assess simultaneous 95% coverage, a conservative Bonferroni adjustment was made to the bootstrap confidence intervals, which used 500 replications: each marginal confidence interval was constructed at the $1 - 0.05/k$ level, where $k = 10$ plans.

Compared to randomized beneficiaries, observational beneficiaries differed across all measured factors: the latter were slightly younger (34.3 vs. 35.5 years old), spent less at baseline ($2796 vs. $3052), were more likely to have a documented sex of female (59% vs. 40%), were less likely to live in Manhattan (13% vs 20%) and more likely to live in Queens (28% vs. 19%), came from different aid groups, and were less likely to be eligible for social security income (2% vs 9%) (Appendix Table B.3 in Appendix B.10). Effect heterogeneity within the randomized data was driven by aid group status, supplemental security income eligibility, and neighborhood effects; within the observational data it was driven by neighborhood effects and receiving aid for children, all of which were imbalanced across randomized and observational beneficiaries, highlighting the need for generalizability approaches.

Overall, across all measured covariates, observational beneficiary characteristics were imbalanced across health plans, and these characteristics were also associated with health care spending, providing empirical evidence that these variables may be confounders. While randomized beneficiaries were not representative of their observational counterparts, there was considerable covariate overlap, as measured by the propensity score for selection into the randomized subset of the data, though overlap was weakest where the observational data was most concentrated (Appendix Figure B.10 in Appendix B.10). Using the conservative overlap hyperparameters $\alpha = 0.01 \times \text{range}(\text{logit}(\pi_S))$ and $\beta = 0.01 \times n_{\text{RCT}}$ resulted in 60% of the target sample within the overlap region. The standardized mean difference in the propensity score for selection was 1.1 standard deviations, which far exceeds 0.25, one proposed threshold indicating large extrapolation (Stuart *et al.*, 2011), and, thus, supportive

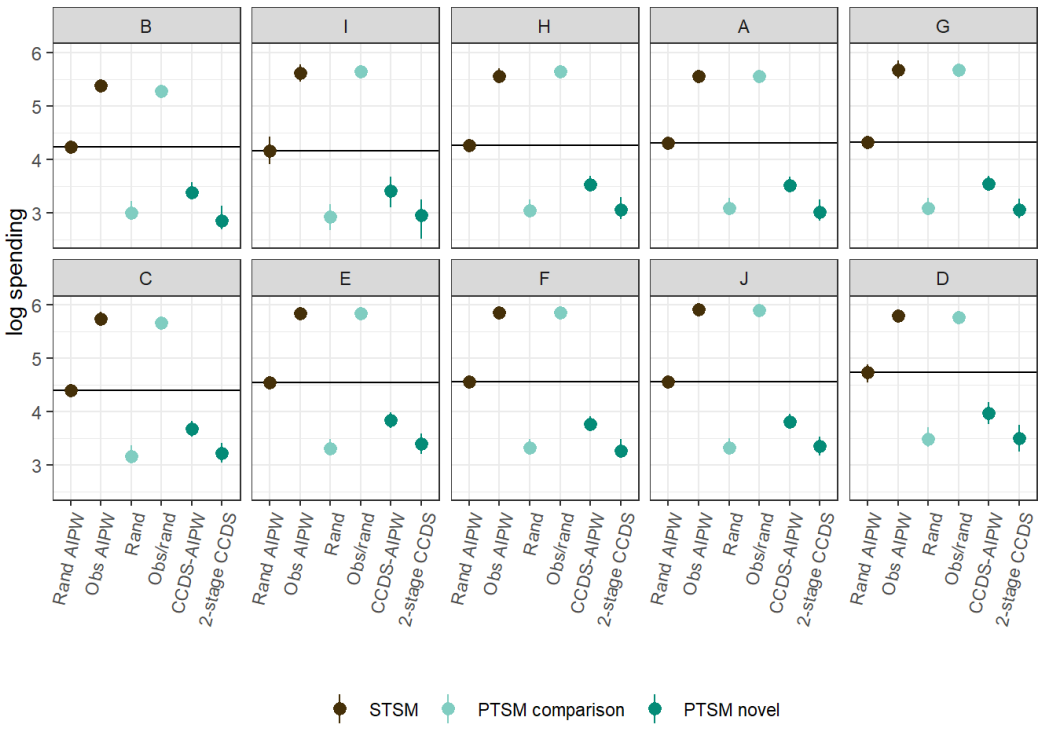



**Figure (2.4)** *STSMs and PTSMs across managed care plans, with 95% multiplicity-adjusted confidence intervals*

of the need for CCDS estimators.

Figure 2.4 presents STSMs for the randomized and observational study populations and PTSMs for the NYC Medicaid target population, including results for two CCDS estimators well suited to this setting. (All estimators are available in Appendix Figure B.11 in Appendix B.10.) Despite higher unadjusted mean spending in the randomized group, causal estimates of STSMs in the observational data were consistently higher than estimates of STSMs in the randomized data across all health plans. This discrepancy reflects both a difference in population characteristics as well as potential unmeasured confounding in the observational data; neither estimate aligned with rand or CCDS estimates of PTSMs, which were consistently lower than randomized and observational STSMs. While the randomized data STSMs showed a difference of 12% between the highest and lowest spending health plans and the observational data STSMs showed a difference of 8%, the rand and CCDS-AIPW estimators

demonstrate a difference of 16% and 17%, respectively, in PTSMs. Thus, study estimates underestimated the true variability in spending between plans.

Given the substantial overlap between randomized and observational covariate distributions, it is unsurprising that CCDS estimates were in a similar range to the rand estimates of PTSMs. However, the double robust CCDS-AIPW estimates were higher than rand estimates (12.5-16.7% difference in log spending) and confidence intervals were non-overlapping for all but plans D and I. CCDS-AIPW also did not show large variability with ensemble regressions, unlike in the simulations. Obs/rand estimates and observational data AIPW STSMs (which largely aligned as the observational data comprised 93% of the data) were widely discrepant from other PTSMs, suggesting a large amount of unmeasured confounding bias in the observational data. Unlike in the simulation, CCDS-IPW confidence intervals were wider than those of other CCDS estimators, which is common to IPW estimators in practice, and also reflects the difficulty of estimating propensities for multiple treatments. While the rand PTSM estimator could provide reasonable estimates in this setting, where there is a fair amount of overlap between randomized and observational data, the CCDS estimators were able to incorporate all data and did not rely on extrapolating spending estimates beyond the support of the randomized data.

## 2.7 Discussion

When observational and randomized data are both available, there is potential to overcome each data type's limitations through their combination. Namely, when some individuals in the target population are not well-represented in the randomized data and the observational data has unmeasured confounding, neither data type alone can successfully generalize to the target population represented by a union of randomized and observational data. This article proposes a class of novel estimators that can surmount positivity of selection assumption violations in the randomized data and unmeasured confounding in the observational data by using common support between the data sources to remove unmeasured confounding bias.

The proposed outcome regression, propensity score, and double robust CCDS estimators

have varying strengths. When the functional forms of the true data generating processes can be approximated by simple linear regressions, the double robust CCDS-AIPW estimator with linear regressions is a suitable default approach for combining randomized and observational data. Even when linear regressions do not capture the full complexity of the data generating process, simulations showed that CCDS-AIPW and CCDS-OR with main terms regressions were able to recover unbiased estimates. However, when fitting more complex regressions, these estimators may lead to unstable bias extrapolations from the overlap region, although we did not see this drawback in our NYC Medicaid data analysis, which had a larger area of overlap. When more complex regression approaches are used, the 2-stage CCDS or CCDS-IPW may also be suitable depending on whether there is more knowledge of the outcome relationship or the propensity for selection and treatment relationships, and whether selection or treatments are rare or multinomial with small probabilities. The 2-stage approach improves performance compared to the CCDS-OR estimator by stabilizing initial estimates to alleviate overfitting to overlap region trends.

In the NYC Medicaid data, there were marked differences between the study and target population causal estimates. Novel and existing generalizability methods helped reconcile these discrepancies by specifying a target population for which inference was desired. There were also significant differences between PTSM rand and obs/rand estimates, showcasing the need to account for both potentially poor extrapolation from the randomized data and potential unmeasured confounding in the observational data. The proposed CCDS estimators provided evidence that the observational data remained subject to unmeasured confounding bias even after adjusting for measured factors.

Our CCDS framework is sensitive to the randomized data regression in the overlap region being an accurate reflection of the truth, as highlighted in the simulation results. When the overlap region is small, the conditional mean relationships estimated from the overlap region may be misspecified, leading to bias and large variability in estimates of unmeasured confounding bias. To assess goodness of fit, investigators can compare estimates to the truth in the randomized data overlap region. Regularization and cross-validation can reduce

chances of overfitting to the data, particularly with more flexible regression approaches. Further practical challenges to applying CCDS estimators in other settings may include that of imperfect covariate correspondence between observational and randomized data sources. Our approach assumes that, after incorporating common covariates, there are no unmeasured outcome determinants (for estimating PTSMs) or effect modifiers (for estimating PATEs) that differ in distribution between between randomized and observational groups. However, if this assumption is violated, CCDS estimators often performed better than using randomized data alone.

Future extensions to the CCDS estimation framework could consider addressing positivity of treatment assignment violations, combining more than two studies (with at least one randomized and one observational), alternative approaches for determining the overlap region that allow for the degree of information borrowing to depend on the similarity of randomized and observational observations, and overlap estimation that does not rely on an estimated propensity score for selection, such as a convex hull approach (King and Zeng, 2006) or estimating common causal support (Hill and Su, 2013). These possible extensions are discussed further in Appendix B.11. Randomized and observational data commonly face multiple challenges beyond those of positivity of selection violation and unmeasured confounding discussed here. These challenges include lack of independence between observations (e.g., clustering), missing data, and measurement error. Methods for addressing such challenges can be combined with our CCDS approaches.

Our CCDS estimators have relevance to many other settings. Positivity of selection violation and unmeasured confounding arise in other studies where the target population is composed of randomized and observational subsets, or more broadly when observational data is being combined with randomized data. For example, in comprehensive cohort studies, patients who refuse randomization are enrolled in a parallel observational study (Lu *et al.*, 2019; Olschewski and Scheurlen, 1985) and when randomized controlled trials are embedded in electronic health record data, the observational data can provide information on patients included in and excluded from the trial (Kibbelaar *et al.*, 2017). Policy evaluation studies

can be combined with observational data from outside the evaluation geography to estimate scale-up impacts (Attanasio *et al.*, 2003; Kern *et al.*, 2016). Across these settings, CCDS estimators can be used to generalize to the target population represented by the union of the randomized and observational data. CCDS could also be applied when randomized data represent the target population but will be combined with observational data to increase power, such as in clinical trials that use a mix of randomized and historical controls (Ghadessi *et al.*, 2020), or when, in the absence of a comprehensive target sample, a combination of randomized and observational studies may more fully represent the target population than either study alone (Prentice *et al.*, 2005; Vaitsiakhovich *et al.*, 2018).

Generalizability methods applied to a specified target population are necessary to obtain unbiased estimates for a policy-relevant population. The internal validity of randomized studies is insufficient to obtain unbiased causal estimates; external validity also needs to be considered. The CCDS estimators presented here provide several approaches for combining randomized with observational data to make inferences that do not rely on extrapolating beyond randomized data support nor on the assumption of unmeasured confounding in the observational data.

# Chapter 3

# Estimating Target Population Average Treatment Effects Among the Treated for a Voluntary Intervention

Irina Degtiar[1], Sherri Rose[2], Mariel Finucane[3]

[1] *Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA*

[2] *Center for Health Policy and Center for Primary Care and Outcomes Research, Stanford University, CA, USA*

[3] *Mathematica Policy Research, Cambridge, MA, USA*

## Abstract

Impact evaluations are often intended to inform future program implementation decisions. However, the context in which a program will be implemented may differ, sometimes substantially, from the context in the evaluation. This difference leads to uncertainty regarding the relevance of evaluation findings to future decisions. Generalizing to future contexts is further challenged when estimating treatment effects among future participants of a voluntary intervention, as future volunteers are not an enumerable population. We present a novel approach for estimating target population average treatment effects among the treated (PATT) by generalizing results from an observational study to target population volunteers (the treated group). Our estimation approach can accommodate flexible outcome regression estimators such as Bayesian Additive Regression Trees (BART) and Bayesian Causal Forests (BCF). BART is a flexible outcome regression that models the response surface as a sum of trees. BCF extends BART by separately regularizing confounding and effect modification components of the outcome regression, which enables full confounding control while shrinking treatment effect heterogeneity. Our generalizability approach incorporates uncertainty regarding target population treatment status into the posterior credible intervals to better-reflect the uncertainty of scaling up a voluntary intervention. In a simulation based on real data, we applied our PATT estimation approach to estimate impacts of a national scale-up of a voluntary health policy model.

## 3.1 Background

Policy impact evaluations seek to guide future policy decisions, such as whether to scale up an intervention at the conclusion of the evaluation study. Much consideration is given to ensuring that impact estimates are internally valid and that all confounders have been measured and adjusted for. Less commonly considered is the external validity of the study, namely, the extent to which findings can hold for future contexts and populations. Study effects may not necessarily hold for a different population when study subjects respond differently to the intervention—when effect modification exists. For example, a national expansion of a Medicare model would not show the same effects as those observed in the evaluation when high-risk patients respond to the intervention differently than low-risk patients, and the proportion of high-risk patients differs between the study population and target population of interest.

Addressing the discrepancy between study and target populations to extend impact results beyond the evaluation at hand requires generalizability and transportability methods. These methods have attracted increasing attention (Degtiar and Rose, 2021), resulting in approaches that make use of outcome regressions such as ordinary least squares (Flores and Mitnik, 2013; Kern *et al.*, 2016) or Bayesian Additive Regression Trees (BART) (Hill, 2011; Green and Kern, 2012; Kern *et al.*, 2016), propensity of selection weighting approaches like Inverse Probability of Participation Weighting (IPPW) (Cole and Stuart, 2010; Correa *et al.*, 2018), and double robust estimators such as the Targeted Maximum Likelihood Estimator (TMLE) (Rudolph and van Der Laan, 2017) and augmented inverse probability of participation weighting (AIPPW) (Dahabreh *et al.*, 2018). However, most of these approaches, with the exception of BART, have to date relied on parametric modeling assumptions. Few existing approaches allow for flexible modeling, which is particularly important when generalizing results from large observational studies with many confounders and effect modifiers.

Policy scalability requires additional considerations novel to the generalizability literature. Inference on the treated population is typically of interest. However, when the intervention is voluntary, the policy has an uncertain target treated population, as it is unclear who

would volunteer for the scale-up. Although there exists a large literature on scale-up implementation considerations (World Health Organization, 2010; Powell *et al.*, 2015; Barker *et al.*, 2016), and several approaches exist for estimating impacts of a policy model scale-up (Attanasio *et al.*, 2003; Flores and Mitnik, 2013; Gechter, 2015), no literature to our knowledge addresses generalizability to a target treated population that is not enumerable (due to uncertainty as to which units would volunteer for the scaled-up intervention in new geographic regions).

To address these shortcomings, we present a novel generalizability approach for estimating target population average treatment effects among the treated (PATT). The approach can accommodate nonparametric outcome regression estimators such as BART and Bayesian Causal Forests (BCF) (Hahn *et al.*, 2020) to extend inference from an observational or randomized study sample to the target treated population while accounting for confounding and effect heterogeneity in a data-driven fashion. BCF has shown superior performance compared to other causal estimators for confounding adjustment (Hahn *et al.*, 2020) and is particularly well-suited for extension to generalizability and transportability settings as it explicitly considers and separately regularizes confounding and effect modification. As a Bayesian estimator, it also allows for incorporation of additional sources of uncertainty into the credible intervals, such as uncertainty around what will drive participation in new geographic regions. Incorporating this source of uncertainty avoids overstating confidence in estimated scale-up effects.

The PATT generalizability estimator first estimates a propensity to volunteer for all units in the target sample, the sample to which the scaled-up intervention will be offered, then estimates impacts for all target sample units from a regression fit to the evaluation sample using BART, BCF, or other estimator. Impacts for the treated sample of volunteers consist of propensity-for-volunteering weighted averages of target sample impact estimates. We apply this novel generalizability estimator to a policy-relevant simulation based on real data to estimate the impact of scaling up a voluntary Medicare model from the evaluation study nationwide.

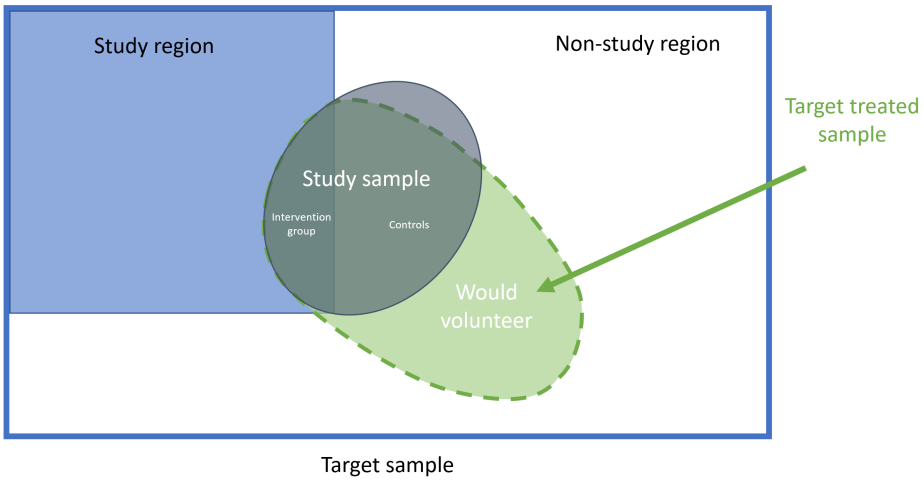



**Figure (3.1)** *Volunteering in study and target samples and regions*

The remainder of this article is organized as follows. Section 2 introduces notation, the estimand, assumptions, and identification. Section 3 presents our novel approach for estimating the PATT and introduces BART, BCF, and other estimator choices. Section 4 assesses our generalizability estimator's performance in a simulation study based on real data. We conclude with a discussion in Section 5.

## 3.2 Notation, assumptions, and causal quantities

### 3.2.1 Data structure, notation, and estimand

The data are generated through the following process, illustrated in Figure 3.1. During the evaluation, the intervention is offered within the study (evaluation) region. Study region units who volunteer to participate in the intervention become part of the study treated population. In the Medicare study, controls were matched to treated units from neighboring regions, though may be chosen in another manner in other studies. In the scale-up, the intervention would be offered to the target population in study and non-study regions. Target population units who volunteer to participate become part of the target treated population. The study region sample, study sample, target sample, and target treated samples are representative samples from their respective populations.

Let $Y$ be the outcome, $S$ be an indicator for being in the study sample, $R$ be an indicator for being in the study region, $V$ be an indicator for volunteering to participate, $A$ be the study treatment, $X$ be the vector of measured covariates, and $n$ be the number of units in the target sample. Capital letters denote random variables while their realizations are denoted in lowercase letters (e.g., $a$, $x$). The observational unit for the target sample is $O_n = (Y, A, X, S, R, V)$ and the observational unit for the target treated sample is $O = (Y, A, X, S, R, V = 1)$. The observational unit for the study region is $O_R = (Y, A, X, S, R = 1, V)$, the observational unit for the study sample is $O_S = (Y, A, X, S = 1, R, V)$ and the observational unit for the study treated sample is $O_A = (Y, A = 1, X, S = 1, R, V = 1)$.

As per the potential outcomes framework, let $Y^1$ be the potential outcome corresponding to participating in the scale-up intervention and $Y^0$ be the potential outcome corresponding to not participating. The estimand of interest is the target population average treatment effect among the treated (PATT): the treatment effect in the target treated population, i.e. among target population units who would volunteer for the intervention: $E(Y^1 - Y^0|V = 1)$.

### 3.2.2 Assumptions

PATT generalizability relies on standard internal validity assumptions, standard external validity assumptions pertaining to the target treated population of volunteers (Stuart *et al.*, 2011; Tipton, 2013a; Degtiar and Rose, 2021), and an additional assumption regarding volunteering:

**Internal validity assumptions:**

1. *Conditional mean exchangeability of treatment assignment:* $E(Y^1 - Y^0|S = 1, X) = E(Y^1 - Y^0|S = 1, A, X)$. The evaluation study was not subject to unmeasured confounding: we adjusted for all variables that risk inducing internal validity bias if not appropriately accounted for.

2. *Positivity of treatment assignment:* $P(X = x|S = 1) > 0 \implies P(A = a|X = x, S = 1) > 0$ with probability 1 for $a = \{0, 1\}$. All units in the study sample would have a positive

probability of being in the intervention group had the intervention been offered to them.

3. *Stable unit treatment value assumption (SUTVA) for treatment assignment:* if $A = a$ then $Y = Y^a$. Units do not impact each other's outcomes and hence there is not nor will be an added benefit nor detriment from being in the same region as an existing intervention participant, and—while participants may have individually made different changes as a result of their intervention participation—the intervention is well-defined for all units.

**External validity assumptions:**

4. *Conditional mean exchangeability of sample selection:* $E(Y^1 - Y^0|V = 1, X) = E(Y^1 - Y^0|S = 1, X)$. There are no unmeasured effect modifiers related to study membership. Thus, new enrollees in the scale-up can be expected to benefit to a similar degree as current participants with similar measured characteristics.

5. *Positivity of sample selection among volunteers:* $P(X = x|V = 1) > 0 \implies P(S = 1|V = 1, X = x) > 0$ with probability 1. The target treated population could have taken part in the current study had it been implemented in their geography.

6. *SUTVA for sample selection:* if $S = s$, and $A = a$ then $Y = Y^a$. Potential outcomes are not a function of how many units are in the intervention, the scale-up intervention will not differ from the evaluation intervention (implementation by unit will remain the same), the study treated population would see similar benefits as those they observed to date, and—if the intervention were terminated—these current study treated participants would revert back to their pre-intervention outcomes.

**Volunteering assumption:**

7. *Equivalent drivers of volunteering:* $P(V = 1|X) = P(V = 1|R = 1, X)$. Which target population units volunteer for the scale-up would be driven by measured characteristics in a similar way as what drove study treated units to participate within study regions.

### 3.2.3 Identification of the estimand

Under the above assumptions, the estimand can be identified as follows:

$$E(Y^1 - Y^0|V=1) = \{E_X[w(X)]\}^{-1} E_X[w(X)\tau(X)]$$

where $w(X) = P(V=1|R=1,X)$ and $\tau(X)$=$E(Y|S=1,A=1,X) - E(Y|S=1,A=0,X)$.

*Proof*:

$$\begin{aligned}
E(Y^1 - Y^0|V=1) &= E_X[E(Y^1|V=1,X) - E(Y^0|V=1,X)|V=1] && (3.1)\\
&= E_X[E(Y^1|S=1,X) - E(Y^0|S=1,X)|V=1] && (3.2)\\
&= E_X[E(Y^1|S=1,A=1,X) - E(Y^0|S=1,A=0,X)|V=1] && (3.3)\\
&= \frac{1}{P(V=1)} E_X[V\{E(Y|S=1,A=1,X) - E(Y|S=1,A=0,X)\}] && (3.4)\\
&= \frac{1}{E_X[P(V=1|X)]} E_X[P(V=1|X)\{E(Y|S=1,A=1,X) - E(Y|S=1,A=0,X)\}] && (3.5)\\
&= \frac{1}{E_X[P(V=1|R=1,X)]} E_X[P(V=1|R=1,X)\{E(Y|S=1,A=1,X) - E(Y|S=1,A=0,X)\}] && (3.6)\\
&= \{E_X[w(X)]\}^{-1} E_X[w(X)\tau(X)] && (3.7)
\end{aligned}$$

Line 3.1 follows from the law of total probability, line 3.2 from conditional sampling exchangeability, line 3.3 from conditional treatment exchangeability, line 3.4 from Bayes rule and SUTVA for treatment assignment and sample selection, line 3.5 from the law of total probability, and line 3.6 from equivalent drivers of volunteering. Positivity of treatment assignment and sample selection among volunteers are needed for the functionals to be well-defined.

## 3.3 Estimating target population average treatment effects among the treated

### 3.3.1 Estimation

The PATT can be estimated by reweighting the target sample to resemble the target treated sample: $\hat{E}(Y^1 - Y^0|V=1) = 1/M \sum_{m=1}^{M} \left\{ [\sum_{i=1}^{n} w^m(X_i)]^{-1} \sum_{i=1}^{n} w^m(X_i)\tau^m(X_i) \right\}$, with $M$ cor-

responding to the number of posterior draws, and $w^m$ and $\tau^m$ representing the $m$-th draw from the posterior distribution of $w$ and $\tau$, respectively. A frequentist approach would estimate $\hat{E}(Y^1 - Y^0 | V = 1) = [\sum_{i=1}^{n} \hat{w}(X_i)]^{-1} \sum_{i=1}^{n} \hat{w}(X_i)\hat{\tau}(X_i)$, with hats denoting point estimates of their respective quantities.

We can similarly estimate the target population conditional average treatment effects among the treated (PCATTs) as $\hat{E}(Y^1 - Y^0 | V = 1, X = x) = 1/M \sum_{m=1}^{M} \left\{ [\sum_{i=1}^{n} w^m(x_i)]^{-1} \sum_{i=1}^{n} w^m(x_i)\tau^m(x_i) \right\}$.

Estimating the PATT requires estimating propensity for volunteering weights and treatment effects for all units in the target sample:

1. *Estimate propensity for volunteering weights, $w$:* Fit a propensity for volunteering regression in the study region, then use it to estimate a posterior distribution of propensities for all units in the target sample.

2. *Estimate treatment effects, $\tau$:* Fit an outcome regression to the study sample, then use it to estimate a posterior distribution of treatment effects for all units in the target sample.

The PATT estimate consists of the mean across posterior draws of propensity-for-volunteering weighted averages of all target sample treatment effects.

### 3.3.2 Incorporating uncertainty with respect to target treated sample membership

To account for the uncertainty in estimating volunteering status when predicting target treated sample impacts, each posterior draw of $\tau$ is multiplied by a different posterior draw of $w$. This sequential approach, though not fully Bayesian, is an unbiased and valid approach for propagating uncertainty in the propensity for volunteering (Zigler and Dominici, 2014; McCandless *et al.*, 2009), given identifiability assumptions. Propensity scores for volunteering should therefore be estimated using a Bayesian model that produces at least as many posterior samples as produced for $\tau$. For a frequentist estimation approach, uncertainty around the

propensity to volunteer can be incorporated using a bootstrap in which both treatment effects and volunteering weights are re-estimated.

### 3.3.3 Bayesian Additive Regression Trees and Bayesian Causal Forests for generalizability

We recommend obtaining $\tau$ estimates with a flexible regression approach such as BART or BCF, which flexibly model the response surface. BART is a Bayesian nonparametric outcome regression (Chipman *et al.*, 2010; Hill, 2011; Green and Kern, 2012; Kern *et al.*, 2016). BART models the outcomes as a sum of binary regression trees with additive error: $Y_i = f(x_i, a_i) + \epsilon_i$ with $f(x_i, a_i) = \sum_{j=1}^{n_{\text{trees}}} g(x_i, a_i, T_j, M_j)$; $\epsilon_i \sim N(0, \sigma^2)$, $T_j$ denotes the tree structure for tree $j$, $M_j$ denotes the bottom node means of tree $j$, $n_{trees}$ is the number of trees, and $\sigma^2$ is the error. A prior is placed on the $g$ tree functions to constrain each tree to be small with $M_j$ near zero (Hill, 2011). However, this prior may create "regularization-induced confounding" by over-shrinking confounding effects (Hahn *et al.*, 2018, 2020).

To overcome the risk of regularization-induced confounding, BCF introduces the propensity score for treatment assignment, $\pi_A(x_i) = P(A_i = 1|x_i)$, as an additional covariate and reparametrizes $f$ to allow for separate priors to be placed on confounding and effect modification (Hahn *et al.*, 2020): $Y_i = \mu(x_i, \pi_A(x_i)) + \tau(x_i)a_i + \epsilon_i$. The function $\mu$ captures the relationship between the outcome and confounders, while $\tau(x_i)a_i$ captures the relationship between the outcome and effect modifiers such that $\tau$ is the treatment effect. Errors may be heteroskedastic (Delannoy *et al.*): $\epsilon_i \sim N(0, {\sigma_i}^2)$.

As with BART, BCF is insensitive to hyperparameter specifications, thus requiring little hyperparameter tuning (the default priors work well across a wide range of settings), and it allows for inference through Bayesian posterior sampling (Hahn *et al.*, 2020). BCF has outperformed other causal estimators at causal inference competitions, such as those held at the Atlantic Causal Inference Conference (Dorie *et al.*, 2019). BCF is particularly well-suited to addressing generalizability as it identifies effect modifiers in a data-driven fashion rather than relying on subjective judgements to estimate heterogeneous treatment effects, it allows

for full confounding control by separately regularizing confounding and effect modification, and it enables incorporation of additional sources of uncertainty in a Bayesian fashion, such as uncertainty with respect to the propensity for volunteering.

In simulations, we explored whether including an estimate of the propensity for being in the study, $\hat{\pi}_S = \hat{P}(S|X)$, as a covariate in the $\tau$ component of the BCF outcome regression would improve finite-sample performance in a similar manner as when including the propensity for treatment as a covariate in the $\mu$ function. Addition of $\hat{\pi}_S$ provides a one-dimensional summary of the association between effect modifiers and selection into the study; the conditional mean exchangeability of sample selection assumption requires that an effect modifier be associated with study membership in order to lead to bias. However, as a practical consideration, including $\hat{\pi}_S$ precludes reusing the same outcome regression already fitted for estimating study impacts. Furthermore, Hahn et al. 2020 found that including the propensity for treatment as an effect modifier can degrade mixing; in simulations, we explore whether including the propensity for being in the study as an effect modifier may likewise hinder mixing.

### 3.3.4 Alternative estimators for $\tau$

The $\tau$ component of the PATT functional can alternatively be estimated through other generalizability approaches such as alternative outcome regression estimators like parametric linear regressions (LR), using IPPW, or using AIPPW. IPPW estimates $\tau(X_i) = Y_i \times w(S_i = 1, A_i = 1, X_i) - Y_i \times w(S_i = 1, A_i = 0, X_i)$ where weights can be normalized for stability $w(S_i = s, A_i = a, X_i) = w^*(S_i = s, A_i = a, X_i) / \sum_{i=1}^{n} w^*(S_i = s, A_i = a, X_i)$ and $w^*(s, a, X_i) = I(S_i = s, A_i = a) / [P(A_i = a|S_i = s, X_i) P(S_i = s|X_i)]$. AIPPW estimates $\tau(X_i) = w(S_i = 1, A_i = 1, X_i)(Y - \hat{E}(Y|S_i = 1, A_i = 1, X_i)) - w(S_i = 1, A_i = 0, X_i)(Y - \hat{E}(Y|S_i = 1, A_i = 0, X_i)) + \hat{E}(Y|S_i = 1, A_i = 1, X_i) - \hat{E}(Y|S_i = 1, A_i = 0, X_i)$. To date, LR, IPPW, and AIPPW estimators for generalizability have relied on parametric regressions which depend on correct model specification for at least one of the regressions (Flores and Mitnik, 2013; Kern *et al.*, 2016; Rudolph and van Der Laan, 2017; Dahabreh *et al.*, 2018). Flexible regression approaches

have demonstrated superior performance to those that rely on parametric assumptions for estimating study population treatment effects (Dorie *et al.*, 2019). The simulation compares these parametric approaches, BART, and BCF for estimating PATTs.

## 3.4 Simulations based on real data

### 3.4.1 Methods

We conducted a simulation to assesses the finite sample performance of our novel PATT estimator for extending inference from an observational study of a voluntary intervention to the target treated sample units who would volunteer to participate in a scale-up. Because estimating treatment effects for a population that is not enumerable (volunteering units) is a novel consideration to the generalizability literature that no existing estimators to our knowledge have addressed, all comparison approaches use our PATT estimation approach, substituting different choices for estimating impacts and weights.

We compared the performance of flexible regressions for estimating $\tau$ and $w$ to using parametric least-squares regressions. Specifically, we compared:

- A BCF outcome regression for $\tau$ that fit a BART propensity model for $w$ ("BCF")
- BCF including the propensity for study selection as an effect modifier with BART propensity models ("BCF-$\pi_S$")
- BART outcome and propensity models ("BART")
- A linear outcome regression that included interactions of each $X$ with $A$ but no higher-order interactions, i.e., between combinations of $X$ with $A$; it estimated the propensity model for $w$ with logistic regression ("LR")
- IPPW with logistic propensity models ("IPPW")
- AIPPW with all regressions estimated using linear and logistic models ("AIPPW")

We generated data to reflect covariate distributions, the outcome generating process, and matched comparison group design of a Medicare study estimating impacts of scaling up a voluntary practice-level Medicare intervention from the evaluation regions nationwide. To do so, we generated simulated study region (n=11,000) and non-study region (n=37,000) data using the conditional distributions $P(Y,A,X,S,R,V) = P(R)P(X|R)P(V|X)P(A,S|R,V,X)P(Y|A,X)$ described in Appendix C.1. Each simulated dataset consisted of approximately 1,000 study treated practices, 4,000 study control practices, and 3,320 practices volunteering from non-study regions for a total of approximately 4,320 volunteering practices (the target treated sample) out of 48,000 nationwide practices (the target sample) (approximately 9% volunteered).

Regressions were fit using practice-level simulated data, and outcome regressions were weighted by the number of beneficiaries in each practice to allow for heteroskedastic errors. We used 2000 replications to ensure a Monte-Carlo standard error of the bias less than 0.1 based on the standard errors of the estimates being less than 4. Inference was conducted at the $\alpha = 0.1$ level, the standard for Medicare evaluations, with empirical 90% uncertainty bounds formed from posterior draws or bootstrap replications. Simulations were run in R using the packages `BCF` (Hahn *et al.*, 2020), `dbarts` Dorie *et al.* (2021), and `MatchIt` (Ho *et al.*, 2011). Each BCF run used 3 chains with 400 posterior samples each after discarding 500 samples as burn-in and thinning by 400 (1200 total posterior samples); each BART run used default settings with 1200 posterior samples.

### 3.4.2 Results

As a result of differences in effect modifier distributions between study and non-study populations, the true study population average treatment effect among the treated (SATT) was \$0.17 (90% sampling variability, i.e., variability in the truth across simulation replications, -2.20 to 2.53) while the true PATT was -\$8.45 (90% sampling variability -9.52 to -7.37). Thus, the intervention showed null impacts in the study but would be cost saving in the target population. All estimators besides IPPW correctly estimated null findings in the study

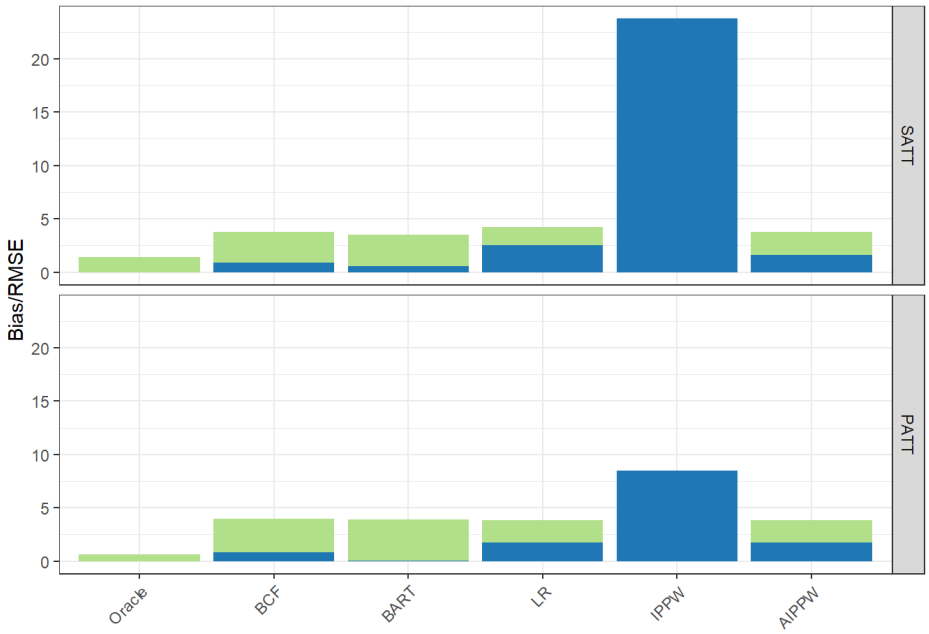



**Figure (3.2)** *Bias and RMSE for SATT and PATT estimates*

*Absolute bias is the darker portion of each bar (in blue); RMSE corresponds to the total bar size.*

population and savings in the target population, though estimators using parametric models overestimated effects by approximately $2 (overestimated the costs in the study treated sample and underestimated the savings in the target treated sample).

Estimators fit with linear and logistic regressions exhibited large bias for estimating both the SATT and the PATT as they were unable to discover the complex confounding and effect heterogeneity relationships observed in the data (Figure 3.2). IPPW's particularly large SATT bias stemmed from the large variability and skewdness in the outcome, since IPPW SATT estimates under treatment simply use the observed outcome for study treated practices. The SATT and PATT were estimated with the smallest bias and RMSE by BART, though all estimators except IPPW exhibited similar RMSE. BART and BCF estimators had the smallest uncertainty bound width (besides IPPW) and attained nominal coverage for the PATT while parametric estimators showed undercoverage (approximately 85% for LR and AIPPW and 0% for IPPW, which is lower than the target of 90%) due to bias in their estimates. All

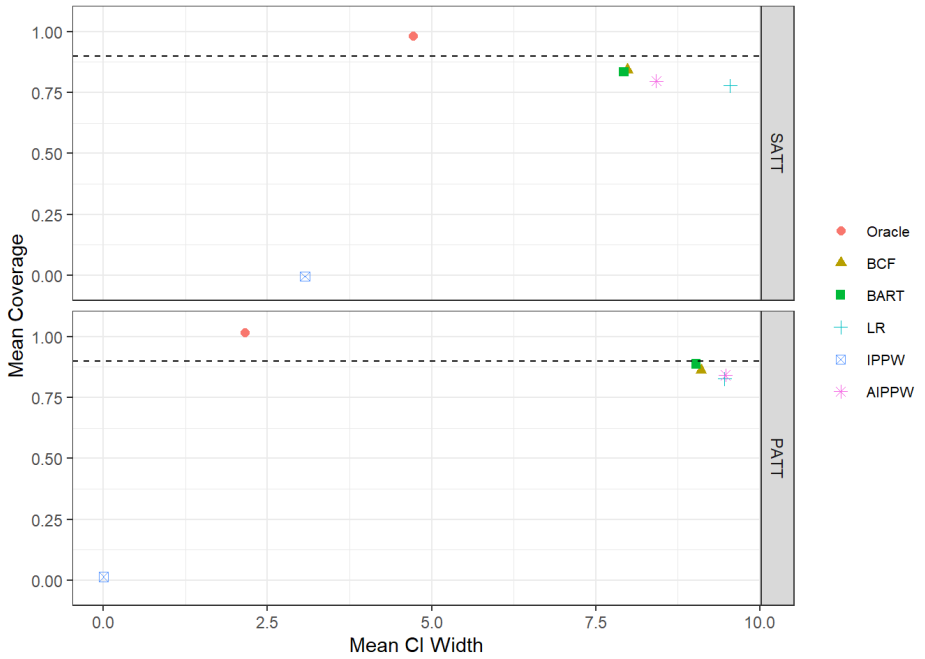



**Figure (3.3)** *Coverage and uncertainty bound width for SATT and PATT estimates*

*The dashed line corresponds to the target coverage of 90%.*

estimators showed undercoverage for estimating SATTs, likely due to unmeasured outcome determinants and model misspecification for estimators using parametric regressions; BCF and BART estimators' coverage was closest to nominal (approximately 85%). Including $\hat{\pi}_S$ as an effect modifier in BCF's regression did not noticeably improve performance (bias and RMSE decreased slightly, by 0.04 and 0.01 respectively for the PATT, a difference smaller than the Monte-Carlo standard error); mixing did not appear hindered by the inclusion of $\hat{\pi}_S$; on average, this inclusion did not impact effective sample sizes nor the potential scale reduction factor.

## 3.5 Discussion

The proposed PATT estimation approach allows for estimating treatment effects among target treated population units that are not enumerable: those who would volunteer for the intervention in a scale-up. Estimating PATTs for a voluntary intervention is a novel

contribution to the generalizability literature that no existing estimator addresses. Our approach does so via a weighted average of treatment effect estimates for all units in the target population to which the intervention would be offered, weighted by the propensity score for volunteering to participate. Posterior credible intervals reflect both uncertainty in treatment effect estimates across target sample practices and uncertainty in target treated sample membership.

In simulations we demonstrated that flexible outcome regression approaches such as BCF and BART improve performance over LR, IPPW, and AIPPW estimators that rely on parametric regressions. BCF and BART estimators not only flexibly adjust for confounding (to ensure internal validity) but also flexibly model effect modification (to ensure external validity). In contrast, parametric regressions were unable to fully account for the complexity of the response surface and so showed bias for both in-sample SATT estimates and out-of-sample PATT estimates. Thus, we demonstrate that flexible regressions improve performance not just for estimating SATTs, as has previously been shown in data analysis competitions (Dorie *et al.*, 2019), but also for estimating PATTs. While BART outperformed BCF for estimating both SATTs and PATTs with our data generating process, across other settings, BCF has demonstrated superior performance to BART (Hahn *et al.*, 2020). In addition to BART and BCF, other flexible approaches for discovering heterogeneous treatment effects could also be used, such as causal forests (Athey and Imbens, 2016; Wager and Athey, 2018), neural network based approaches (Johansson *et al.*, 2018; Shalit *et al.*, 2017), Gaussian process based approaches (Alaa and van der Schaar, 2017, 2018), and ensembles (Grimmer *et al.*, 2017; Lee *et al.*, 2020). In contrast to these estimators, BART and BCF offer uncertainty bounds based on the posterior, ability to easily incorporate other sources of uncertainty due to being Bayesian estimators, and insensitivity to hyperparameter tuning.

A limitation of many flexible modeling approaches like BCF and BART is their computational burden. BCF is currently impractical to fit to data involving millions of observations, rather than thousands, and thus is unworkable for patient-level analyses of health policy interventions in lieu of the practice-level analyses we conducted in our simulations. Even

with practice-level data, BCF took approximately 20 minutes to fit and BART approximately 8 minutes while linear regressions took seconds (albeit their bootstrap took around 11 minutes, on an Intel(R) Xeon(R) CPU E7-8895 v2 @ 2.80GHz processor). Work by Hahn *et al.* (2020) on warm-start BCF, by Pratola *et al.* (2014) on single program multiple data parallel computation for BART, by He *et al.* (2019) on XBART, and others is ongoing to improve the computational efficiency of these methods. BCF's practicality can be further enhanced by extensions to available software to allow for correlated data analysis (Yeager *et al.*, 2019).

The presented estimator assesses the impact of scaling up an intervention such as a policy model as it was offered in the study. It therefore does not capture changes to the intervention under scale-up nor does it account for changes to the setting such as the political landscape. The estimator furthermore relies on identifiability assumptions that preclude unmeasured confounding and unmeasured effect modification, spillover effects, and different drivers of participation between study and non-study regions. To the extent that a prior distribution can be placed on these factors, there is room to incorporate these sources of uncertainty into the credible intervals. Otherwise, sensitivity analyses can assess the impact of such considerations on PATT estimates.

The generalizability estimator presented here can also be applied to identify alternative feasible target populations for a scale-up, such as populations expected to benefit most from the intervention, defined by values of key effect modifiers. Estimates from such targeting approaches can inform future policy models and model expansions, particularly when a broader scale-up is estimated to be ineffective. By estimating treatment effects for the target treated population of interest, our generalizability estimator provides impacts in a policy-relevant population to better-guide future policy decisions.

# References


Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R. and Stuart, E. A. (2019). Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive behaviors*, **94**, 124–132.

Alaa, A. and van der Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In J. Dy and A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol. 80, pp. 129–138.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. *CoRR*, **abs/1704.02801**.

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly journal of economics*, **130** (3), 1117–1166.

— and Mullainathan, S. (2012). External validity and partner selection bias. *National Bureau of Economic Research Working Paper Series*, **18373**, 53.

Andrews, I. and Oster, E. (2017). *Weighting for External Validity*. Tech. Rep. w23826, National Bureau of Economic Research, Cambridge, MA.

Angrist, J. D. and Fernández-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In *Advances in Economics and Econometrics*, Cambridge University Press, pp. 401–434.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113** (27), 7353–7360.

Attanasio, O., Meghir, C. and Szekely, M. (2003). Using randomised experiments and structural models for 'scaling up': Evidence from the PROGRESA evaluation. *IFS Working Paper*, **EWP03/05**.

Baker, R., Brick, J. M., Gotway Crawford, C. A., Terhanian, G., Langer, G., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., Tourangeau, R., Valliant, R. and Rivers, D. (2013). Summary report of the aapor task force on non-probability sampling. *Journal of survey statistics and methodology*, **1** (2), 90–136.

Bareinboim, E. and Pearl, J. (2014). Transportability from multiple environments with limited experiments: Completeness results. *Advances in Neural Information Processing Systems 27*, pp. 280–288.

— and — (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, **113** (27), 7345–7352.

— and Tian, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, AAAI Press, pp. 3475–3481.

—, — and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, AAAI Press, pp. 2410–2416.

Barker, P. M., Reid, A. and Schall, M. W. (2016). A framework for scaling up health interventions: Lessons from large-scale improvement initiatives in Africa. *Implementation Science*, **11** (1), 12.

Begg, C. B. (1992). Cross design synthesis: A new strategy for medical effectiveness research. united states general accounting office, (GA0/PEMD-92-18). *Statistics in Medicine*, **11** (12), 1627–1628.

Bell, S. H., Olsen, R. B., Orr, L. L. and Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, **38** (2), 318–335.

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M. and RECORD Working Committee (2015). The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLOS Medicine*, **12** (10), e1001885.

Bennett, M., Vielma, J. P. and Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *Journal of computational and graphical statistics*, **29** (4), 744–757.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. P. A. and Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, **23** (5), 749–767.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J. and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A, Statistics in society*, **181** (4), 1193–1209.

Burchett, H., Umoquit, M. and Dobrow, M. (2011). How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of health services research & policy*, **16** (4), 238–244.

Cahan, A., Cahan, S. and Cimino, J. J. (2017). Computer-aided assessment of the generalizability of clinical trial results. *International Journal of Medical Informatics*, **99**, 60–66.

Centers for Medicare & Medicaid Services (2020). Medicaid Facts and Figures | CMS. https://www.cms.gov/newsroom/fact-sheets/medicaid-facts-and-figures.

Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, **10** (3), 646–669.

Chen, C. and Wong, R. (2018). *Black Patients Miss out on Promising Cancer Drugs*.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K. and Ghassemi, M. (2020). Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*.

Chen, S., Tian, L., Cai, T. and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, **73** (4), 1199–1209.

Chen, Z. and Kaizar, E. (2017). On variance estimation for generalizing from a trial to a target population. *arXiv:1704.07789 [stat]*.

Chipman, H. A., George, E. I. and McCulloch, R. (2007). Bayesian ensemble learning. In *Advances in Neural Information Processing Systems 19 - Proceedings of the 2006 Conference*, pp. 265–272.

—, — and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, **4** (1), 266–298.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, **172** (1), 107–115.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J. and Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: A review. *arXiv:2011.08047 [stat]*.

Correa, J. D. and Bareinboim, E. (2017). Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, AAAI Press, pp. 3740–3746.

—, Tian, J. and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. In *AAAI*.

Cronbach, L. J. and Shapiro, K. (1982). *Designing Evaluations of Educational and Social Programs*. A Joint Publication in the Jossey-Bass Series in Social and Behavioral Science & in Higher Education, San Francisco: Jossey-Bass, 1st edn.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, **90** (3), 389–405.

Dahabreh, I., Robertson, S., Stuart, E. and Hernan, M. (2017). Extending inferences from randomized participants to all eligible individuals using trials nested within cohort studies. *arXiv:1709.04589 [stat]*.

Dahabreh, I. J., Hernan, M. A., Robertson, S. E., Buchanan, A. and Steingrimsson, J. A. (2019a). Generalizing trial findings using nested trial designs with sub-sampling of non-randomized individuals.

—, Robertson, S. E., Petito, L. C., Hernán, M. A. and Steingrimsson, J. A. (2019b). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population.

—, —, Steingrimsson, J. A., Stuart, E. A. and Hernan, M. A. (2018). Extending inferences from a randomized trial to a new target population. *arXiv:1805.00550 [stat]*.

—, —, Tchetgen, E. J., Stuart, E. A. and Hernán, M. A. (2019c). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, **75** (2), 685–694.

—, Robins, J. M., Haneuse, S. J.-P. A., Saeed, I., Robertson, S. E., Stuart, E. A. and Hernán, M. A. (2019d). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population.

Davis, K. (1988). The comprehensive cohort study: The use of registry data to confirm and extend a randomized trial. *Recent results in cancer research*, **111**, 138.

Degtiar, I. and Rose, S. (2021). A Review of Generalizability and Transportability. *arXiv:2102.11904 [stat]*.

Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A. and Vandenbroucke, J. P. (2010). How to assess the external validity of therapeutic trials: A conceptual approach. *International Journal of Epidemiology*, **39** (1), 89–94.

Delannoy, C., Vollmer Forrow, L. and Finucane, M. (). Bayesian Causal Forests: A Data-Driven Approach to Subgroup Analysis.

Ding, P., Feller, A. and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **78** (3), 655–671.

Dong, N., Stuart, E. A., Lenis, D. and Quynh Nguyen, T. (2020). Using propensity score analysis of survey data to estimate population average treatment effects: A case study comparing different methods. *Evaluation review*, **44** (1), 84–108.

Dorie, V., Chipman, H., McCulloch, R., Dadgar, A., Draheim, G. U., Bosmans, M., Tournayre, C., Petch, M., Valle, R. d. L., Johnson, S. G., Frigo, M., Zaitseff, J., Veldhuizen, T., Maisonobe, L., Pakin, S. and G., D. R. (2021). Dbarts: Discrete Bayesian Additive Regression Trees Sampler.

—, Hill, J., Shalit, U., Scott, M. and Cervone, D. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, **34** (1), 43–68.

Eddy, D. (1989). The confidence profile method: A bayesian method for assessing health technologies. *Operations Research*, **37** (2), 210–228.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton: CRC Press LLC.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical science*, **32** (2), 249–264.

Fang, A. (2017). *10 Things to Know about Heterogeneous Treatment Effects*.

Flores, C. A. and Mitnik, O. A. (2013). Comparing treatments across labor markets: An assessment of nonexperimental multiple-treatment strategies. *The Review of Economics and Statistics*, **95** (5), 1691–1707.

Ford, I. and Norrie, J. (2016). Pragmatic trials. *New England Journal of Medicine*, **375** (5), 454–463.

Frangakis, C. (2009). The calibration of treatment effects from clinical trials to target populations. *Clinical Trials: Journal of the Society for Clinical Trials*, **6** (2), 136–140.

Gabler, N. B., Duan, N., Liao, D., Elmore, J. G., Ganiats, T. G. and Kravitz, R. L. (2009). Dealing with heterogeneity of treatment effects: Is the literature up to the challenge? *Trials*, **10** (1), 43–43.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41** (2), 361.

Gechter, M. (2015). Generalizing the results from social experiments: Theory and evidence from mexico and india. *Department of Economics, Pennsylvania State University*, **Unpublished manuscript**, 50.

Gelman, A. and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, (23), 127–135.

Geruso, M., Layton, T. J. and Wallace, J. (2020). *Are All Managed Care Plans Created Equal? Evidence from Random Plan Assignment in Medicaid*. Tech. Rep. w27762, National Bureau of Economic Research.

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., Mei, C., Zhang, L., Deng, C. Q. and Beckman, R. A. (2020). A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet Journal of Rare Diseases*, **15** (1), 69.

Glauner, P., Migliosi, A., Meira, J., Valtchev, P., State, R. and Bettinger, F. (2017). Is big data sufficient for a reliable detection of non-technical losses? *arXiv:1702.03767 [cs]*.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, **76** (3), 491–511.

Green, L. W. and Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Evaluation & the health professions*, **29** (1), 126–153.

Greenhouse, Kelleher, K., Seltman, H. and Gardner, W. (2008). Generalizing from clinical trial data: A case study. the risk of suicidality among pediatric antidepressant users. *Statistics in Medicine*, **27** (11), 1801–13.

Greenhouse, J. B., Kaizar, E. E., Anderson, H. D., Bridge, J. A., Libby, A. M., Valuck, R. and Kelleher, K. J. (2017). Combining information from multiple data sources: An introduction to cross-design synthesis with a case study. In *Methods in Comparative Effectiveness Research*, Chapman and Hall/CRC, pp. 223–246.

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal Of The Royal Statistical Society Series A*, **168**, 267–291.

Grimmer, J., Messing, S. and Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, **25** (4), 413–434.

Gunter, L., Zhu, J. and Murphy, S. (2011). Variable selection for qualitative interactions. *Statistical Methodology*, **8** (1), 42–55.

Hahn, P. R., Carvalho, C. M., Puelz, D. and He, J. (2018). Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, **13** (1).

—, Murray, J. S. and Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, **15** (3), 965–1056.

Haneuse, S. (2016). Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical care*, **54** (4), e23–e29.

—, Schildcrout, J., Crane, P., Sonnen, J., Breitner, J. and Larson, E. (2009). Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*, **32** (3), 229–239.

Hartman, E., Grieve, R., Ramsahai, R. and Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178** (3), 757–778.

He, J., Yalov, S. and Hahn, P. R. (2019). XBART: Accelerated Bayesian Additive Regression Trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1130–1138.

He, Z., Ryan, P., Hoxha, J., Wang, S., Carini, S., Sim, I. and Weng, C. (2016). Multivariate analysis of the population representativeness of related clinical studies. *Journal of biomedical informatics*, **60**, 66–76.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, **47** (1), 153–161.

Henderson, N. C., Varadhan, R. and Weiss, C. O. (2017). Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogenous: Part II. application and external validation. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **3** (1-2), 7–20.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E. and Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, **19** (6), 766–779.

Hill, J. and Su, Y.-S. (2013). Assessing Lack of Common Support in Causal Inference Using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive Outcomes. *The annals of applied statistics*, **7** (3), 1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20** (1), 217–240.

Ho, D., Imai, K., King, G. and Stuart, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, **42** (1), 1–28.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47** (260), 663–685.

Hotz, V. J., Imbens, G. W. and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of econometrics*, **125** (1), 241–270.

Imai, K., King, G. and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171** (2), 481–502.

Johansson, F. D., Kallus, N., Shalit, U. and Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv:1802.08598 [stat]*.

Josey, K. P., Berkowitz, S. A., Ghosh, D. and Raghavan, S. (2020a). Transporting experimental results with entropy balancing.

—, Yang, F., Ghosh, D. and Raghavan, S. (2020b). A calibration approach to transportability with observational data.

Kaizar, E. E. (2011). Estimating treatment effect via simple cross design synthesis. *Statistics in Medicine*, **30** (25), 2986–3009.

— (2015). Incorporating both randomized and observational data into a single analysis. *Annual Review of Statistics and Its Application*, **2** (1), 49–72.

Kallus, N., Puli, A. M. and Shalit, U. (2018). Removing hidden confounding by experimental grounding. *arXiv:1810.11646 [cs, stat]*.

Kennedy, L. and Gelman, A. (2019). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *arXiv:1906.11323 [stat]*.

Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S. and Johnston, J. (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*, **16** (1), 495–495.

Kern, H. L., Stuart, E. A., Hill, J. and Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, **9** (1), 103–127.

Kibbelaar, R. E., Oortgiesen, B. E., van der Wal-Oost, A. M., Boslooper, K., Coebergh, J. W., Veeger, N. J. G. M., Joosten, P., Storm, H., van Roon, E. N. and Hoogendoorn, M. (2017). Bridging the gap between the randomised clinical trial world and the real world by combination of population-based registry and electronic health record data: A case study in haemato-oncology. *European Journal of Cancer*, **86**, 178–185.

Kim, J. K., Park, S., Chen, Y. and Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation.

King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political analysis*, **14** (2), 131–159.

Lee, K., Bargagli-Stoffi, F. J. and Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G. and Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, **28** (4), 553–561.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011). Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations: Subgroup identification based on differential effect search (SIDES). *Statistics in Medicine*, **30** (21), 2601–21.

Lipman, E., Deke, J. and Finucane, M. (). Bayesian Interpretation of Cluster Robust Subgroup Impact Estimates: The Best of Both Worlds.

Lu, Y., Scharfstein, D. O., Brooks, M. M., Quach, K. and Kennedy, E. H. (2019). Causal inference for comprehensive cohort studies. *arXiv:1910.03531 [stat.ME]*.

Luedtke, A., Carone, M. and van der Laan, M. J. (2019). An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, **81** (1), 75–99.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, **23** (19), 2937–60.

Marcus, S. (1997). Assessing non-consent bias with parallel randomized and nonrandomized clinical trials. *Journal Of Clinical Epidemiology*, **50** (7), 823–828.

McCandless, L. C., Gustafson, P. and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, **28** (1), 94–112.

Medicaid (2020). *Managed Care in New York*. Tech. rep.

Miettinen, O. S. (1972). Standardization of risk ratios. *American Journal of Epidemiology*, **96** (6), 383–388.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, **45** (1), 521–530.

Nethery, R. C., Mealli, F. and Dominici, F. (2018). Estimating Population Average Causal Effects in the Presence of Non-Overlap: The Effect of Natural Gas Compressor Station Exposure on Cancer Mortality. *arXiv:1805.09736 [stat]*.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of statistical planning and inference*, **129** (1-2), 405–426.

Nguyen, T. Q., Ackerman, B., Schmid, I., Cole, S. R. and Stuart, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLOS ONE*, **13** (12), e0208795.

—, Ebnesajjad, C., Cole, S. R. and Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Annals of Applied Statistics*, **11** (1), 225–247.

Nie, L., Zhang, Z., Rubin, D. and Chu, J. (2013). Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. *The Annals of Applied Statistics*, **7** (3), 1796–1813.

O' Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63** (2), 195–210.

Olschewski, M. and Scheurlen, H. (1985). Comprehensive Cohort Study: An alternative to randomized consent design in a breast preservation trial. *Methods of Information in Medicine*, **24** (3), 131–134.

Olsen, R. B., Orr, L. L., Bell, S. H. and Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively: External validity in policy evaluations. *Journal of policy analysis and management*, **32** (1), 107–121.

Pan, Q. and Schaubel, D. E. (2008). Proportional hazards models based on biased samples and estimated selection probabilities. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **36** (1), 111–127.

— and — (2009). Evaluating bias correction in weighted proportional hazards regression. *Lifetime Data Analysis*, **15** (1), 120–146.

Park, D. K., Gelman, A. and Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, (12), 375–385.

Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge University Press.

— (2015). Generalizing experimental findings. *Journal of Causal Inference*, **3** (2).

— and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, Vancouver, BC, Canada: IEEE, pp. 540–547.

— and — (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, **29** (4), 579–595.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R. and Welton, N. J. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical decision making*, **38** (2), 200–211.

Polley, E., LeDell, E., Kennedy, C., Lendle, S. and van der Laan, M. (2019). SuperLearner: Super Learner Prediction.

Pool, I., Abelson, R. and Popkin, S. (1964). *Candidates, Issues and Strategies; a Computer Simulation of the 1960 Presidential Election*. Massachusetts Institute of Technology Press.

Powell, B. J., Waltz, T. J., Chinman, M. J., Damschroder, L. J., Smith, J. L., Matthieu, M. M., Proctor, E. K. and Kirchner, J. E. (2015). A refined compilation of implementation strategies: Results from the Expert Recommendations for Implementing Change (ERIC) project. *Implementation science: IS*, **10**, 21.

Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R. and Rust, W. N. (2014). Parallel Bayesian Additive Regression Trees. *Journal of Computational and Graphical Statistics*, **23** (3), 830–852.

Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., Limacher, M. and Wactawski-Wende, J. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *American Journal of Epidemiology*, **162** (5), 404–414.

—, Langer, R. D., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G. L., Barad, D., Curb, J. D., Kotchen, J., Kuller, L., Limacher, M. and Wactawski-Wende, J. (2006). Combined analysis of Women's Health Initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *American journal of epidemiology*, **163** (7), 589–599.

Prevost, T. C., Abrams, K. R. and Jones, D. R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine*, **19** (24), 3359–3376.

Qian, M., Chakraborty, B. and Maiti, R. (2019). A sequential significance test for treatment by covariate interactions. *arXiv:1901.08738 [stat]*.

Robins, J. M., Hernán, M. A. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, **11** (5), 550–560.

Rose, S. and Normand, S.-L. (2019). Double robust estimation for multiple unordered treatments and clustered observations: Evaluating drug-eluting coronary artery stents. *Biometrics*, **75** (1), 289–296.

Rosenman, E., Basse, G., Owen, A. and Baiocchi, M. (2020). Combining observational and experimental datasets using shrinkage estimators. *arXiv: 2002.06708 [stat.ME]*.

—, Owen, A. B., Baiocchi, M. and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv:1804.07863 [stat]*.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *The Lancet*, **365** (9453), 82–93.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66** (5), 688–701.

Rudolph, K. and van Der Laan, M. (2017). Robust estimation of encouragement design intervention effects transported across sites. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, **79** (5), 1509–1525.

Schmid, I., Rudolph, K. E., Nguyen, T. Q., Hong, H., Seamans, M. J., Ackerman, B. and Stuart, E. A. (2020). Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations. *Communications in statistics. Simulation and computation*, **ahead-of-print** (ahead-of-print), 1–23.

Schulz, K. F., Altman, D. G. and Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, **340** (mar23 1), c332–c332.

Schwartz, D. and Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, **20** (8), 637–648.

Sen, A., Chakrabarti, S., Goldstein, A., Wang, S., Ryan, P. B. and Weng, C. (2016). GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. *Journal of Biomedical Informatics*, **63**, 325–336.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shalit, U., Johansson, F. D. and Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In D. Precup and Y. W. Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, PMLR, *Proceedings of Machine Learning Research*, vol. 70, pp. 3076–3085.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R. and Mulani, P. M. (2010). Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*, **28** (10), 935–945.

Simon, R. (1982). Patient subsets and variation in therapeutic efficacy. *British Journal of Clinical Pharmacology*, **14** (4), 473–482.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25** (1), 1–21.

—, Ackerman, B. and Westreich, D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, **28** (5), 532–537.

—, Bradshaw, C. P. and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, **16** (3), 475–485.

—, Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials: Use of propensity scores to assess generalizability. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174** (2), 369–386.

Su, X., Tsai, C., Wang, H., Nickerson, D. and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal Of Machine Learning Research*, **10**, 141–158.

—, Zhou, T., Yan, X., Fan, J. and Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*, **4** (1).

Tian, L., Alizadeh, A. A., Gentles, A. J. and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, **109** (508), 1517–1532.

Tipton, E. (2013a). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, **38** (3), 239–266.

— (2013b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, **37** (2), 109–139.

— (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, **39** (6), 478–501.

—, Hallberg, K., Hedges, L. V. and Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, **41** (5), 472–505.

—, Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K. and Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, **7** (1), 114–135.

— and Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, **47** (8), 516–524.

— and Peck, L. R. (2017). A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review*, **41** (4), 326–356.

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S. and Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **172** (1), 21–47.

Vaitsiakhovich, T., Filonenko, A., Lynen, R., Endrikat, J. and Gerlinger, C. (2018). Cross design analysis of randomized and observational data - application to continuation rates for a contraceptive intra uterine device containing Levonorgestrel in adolescents and adults. *BMC women's health*, **18** (1), 180–180.

Van der Laan, M. J., Laan, M. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. Springer Series in Statistics, New York, NY: Springer New York.

Varadhan, R., Henderson, N. C. and Weiss, C. O. (2016). Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogeneous: Part i. methodology. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **2** (3-4), 112–126.

Verde, P. E. (2019). The hierarchical metaregression approach and learning from clinical evidence. *Biometrical Journal*, **61** (3), 535–557.

— and Ohmann, C. (2015). Combining randomized and non-randomized evidence in clinical research: A review of methods and applications: Combining randomized and non-randomized evidence. *Research Synthesis Methods*, **6** (1), 45–62.

—, —, Morbach, S. and Icks, A. (2016). Bayesian evidence synthesis for exploring generalizability of treatment effects: A case study of combining randomized and non-randomized results in diabetes: Bayesian evidence synthesis for exploring generalizability of treatment effects: A case study of combining randomized and non-randomized results in di. *Statistics in Medicine*, **35** (10), 1654–1675.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C. and Vandenbroucke, J. P. (2008). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, **61** (4), 344–349.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, **113** (523), 1228–1242.

Weisberg, H. I., Hayden, V. C. and Pontes, V. P. (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clinical Trials*, **6** (2), 109–18.

Weiss, C. O., Segal, J. B. and Varadhan, R. (2012). Assessing the applicability of trial evidence to a target sample in the presence of heterogeneity of treatment effect: APPLICABILITY OF TREATMENT EFFECTS. *Pharmacoepidemiology and Drug Safety*, **21**, 121–129.

Weng, C., Li, Y., Ryan, P., Zhang, Y., Liu, F., Gao, J., Bigger, J. and Hripcsak, G. (2014). A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Applied clinical informatics*, **5** (2), 463–479.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E. and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, **186** (8), 1010–1014.

World Health Organization (2010). Nine steps for developing a scaling-up strategy. https://implementationscience.biomedcentral.com/articles/10.1186/s13012-016-0374-x.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L. and Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, **573** (7774), 364–369.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, **109** (505), 95–107.

# Appendix A

# Appendix to Chapter 1

## A.1 Summary of methods that only require summary-level data

Without access to individual patient data in the study and/or target samples, investigators will be constrained as to the estimators available to them. The following estimators can be applied in this setting. Investigators should strive to maximally use the available data and hence use methods that incorporate individual-level data where they are available.

**Summary-level data for both study (covariate and outcome) and target samples (covariate).** Post-stratification (Miettinen, 1972; Prentice *et al.*, 2005) only requires joint distributions or cell counts for each stratum. Using only study and target sample means, one could also apply outcome regressions that are linear in their predictors.

**Summary-level outcome data for both study and target samples.** Bias-adjusted meta-analysis approaches by Turner *et al.* (2009) and Greenland (2005) require summary-level study outcome data with estimates of bias for each study. When that summary-level data are stratified by effect modifiers, one can use approaches by Eddy (1989) and Prevost *et al.* (2000). If summary-level study data are stratified by participants included vs. excluded from the study, cross-design synthesis can be used (Begg, 1992; Kaizar, 2011).

**Summary-level covariate and outcome data in the study, individual-level covariate and outcome data in the target sample.** With summary-level study and individual-level target sample data, one can use hierarchical Bayesian evidence synthesis (Verde *et al.*, 2016; Verde, 2019).

**Individual-level covariate and outcome data in the study, summary-level covariate data in the target sample.** With individual-level study and summary-level target data, one can use matching with reweighting (e.g., Hartman *et al.* (2015)), or Signorovitch *et al.* (2010) or Phillippo *et al.* (2018)'s propensity and outcome regression approaches. When joint distributions of summary-level target sample data are available, one can use IPPW (Cole and Stuart, 2010; Westreich *et al.*, 2017).

# Appendix B

# Appendix to Chapter 2

## B.1 Derivation of Assumption 1b

To overcome violations of Assumptions 1 (mean conditional treatment exchangeability, or no unmeasured confounding) and 5 (positivity of study selection), we can leverage information from the combination of randomized and observational data.

To tackle the unmeasured confounding bias, let's begin by characterizing the conditional bias in the observational group: $b(a,\boldsymbol{x}) \equiv E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=0,\boldsymbol{X}=\boldsymbol{x})$. The conditional bias corresponds to the average difference in potential outcomes between observational group individuals on intervention $a$ vs. marginally, conditioning on measured covariates $\boldsymbol{X}=\boldsymbol{x}$. We could alternatively have defined conditional bias relative to a specific alternative intervention, $E(Y^a|S=0,A=a',\boldsymbol{X}=\boldsymbol{x})$, or relative to all other interventions, $E(Y^a|S=0,A\neq a,\boldsymbol{X}=\boldsymbol{x})$; the same principles hold. Mean conditional treatment exchangeability holds if and only if $b(a,\boldsymbol{X})=0$ for all $a\in\mathcal{A}$.

By randomization, mean conditional treatment exchangeability holds for the randomized group, hence $E(Y^a|S=1,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=1,\boldsymbol{X}=\boldsymbol{x}) = 0$. We therefore have that:

$$\begin{aligned} b(a,\boldsymbol{x}) = {} & E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=1,A=a,\boldsymbol{X}=\boldsymbol{x}) \\ & - \left[E(Y^a|S=0,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=1,\boldsymbol{X}=\boldsymbol{x})\right] \end{aligned}$$

By mean conditional exchangeability for study selection, $E(Y^a|S=1,\boldsymbol{X}) = E(Y^a|S=0,\boldsymbol{X})$ and thus $b(a,\boldsymbol{x}) = E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=1,A=a,\boldsymbol{X}=\boldsymbol{x})$.

However, overlapping support between randomized and observational groups only exists in $\mathcal{R}_{\text{overlap}}$, hence $E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=1,A=a,\boldsymbol{X}=\boldsymbol{x})$ can only be identified in $\mathcal{R}_{\text{overlap}}$ without further assumptions to warrant the extrapolation. One extrapolation approach would be to directly extrapolate from the randomized group to obtain potential outcomes in regions of non-support (see the rand estimator in Section 2.4.6 for an estimation strategy based on this approach), or to extrapolate for the purposes of estimating bias in regions of non-support (see the 2-stage WD estimator in Appendix B.6), but estimation relying on these strategies is sensitive to parametric assumptions needed to extrapolate beyond the randomized data's support. We instead make an alternative assumption, Assumption 1b: $b(a,\boldsymbol{x}) = b(a,\boldsymbol{x}|R_{\text{overlap}}=1)$; namely, that the same conditional bias relationship that holds in the region of overlap also holds in the broader support of the observational group. When estimating PTSMs, more weakly, the constant conditional bias assumption must hold in expectation over the $\boldsymbol{X}$ distribution in the observational data: $E_{\boldsymbol{X}}\Big[b(a,\boldsymbol{x})\Big|S=0\Big] = E_{\boldsymbol{X}}\Big[b(a,\boldsymbol{x}|R_{\text{overlap}}=1)\Big|S=0\Big]$.

The mean constant conditional bias assumption can also be restated with respect to the unmeasured confounders that are implicitly being integrated over. Assumption 1b states that, in expectation, the bias when integrating over the distribution of unmeasured confounders in $\mathcal{R}_{\text{overlap}}$ is equivalent to the bias when integrating over the distribution of unmeasured confounders in $\mathcal{R}_{\text{obs}}$. Namely, with $\boldsymbol{U}$ corresponding to unmeasured confounders, the mean constant bias assumption can be written as:

$$
\begin{aligned}
&E_{\boldsymbol{X}}\Big\{E_{\boldsymbol{U}}\Big[E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X},\boldsymbol{U})|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}\Big] \\
&\qquad - E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big|S=0\Big\} \\
&= E_{\boldsymbol{X}}\Big\{E_{\boldsymbol{U}}\Big[E(Y^a|S=0,A=a,\boldsymbol{X},\boldsymbol{U})|S=0,A=a,\boldsymbol{X}\Big] \\
&\qquad - E(Y^a|S=1,A=a,\boldsymbol{X})\Big|S=0\Big\}
\end{aligned}
$$

for all $a \in \mathcal{A}$.

## B.2 Sensitivity analysis bounds

Making no constant conditional bias assumptions, we arrive at the following functional of the observed data and potential outcomes:

$$\begin{aligned} E(Y^a) =& E_{\boldsymbol{X}}[E(Y|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &+ E_{\boldsymbol{X}}[E(Y|S=0,A=a,\boldsymbol{X}) - b'(a,x)P(A\neq a|S=0,\boldsymbol{X})|S=0]P(S=0) \end{aligned} \tag{B.1}$$

where $b'(a,\boldsymbol{x}) = E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=0,A\neq a,\boldsymbol{X}=\boldsymbol{x})$

*Proof for B.1:* Using the law of iterated expectations, no unmeasured confounding in the randomized group, SUTVA assumptions, and positivity assumptions, we obtain the following:

$$\begin{aligned} E(Y^a) &= E(Y^a|S=1)P(S=1) + E(Y^a|S=0)P(S=0) \\ &= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &\quad + E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,\boldsymbol{X})P(A=a|S=0,\boldsymbol{X}) \\ &\quad + E(Y^a|S=0,A\neq a,\boldsymbol{X})P(A\neq a|S=0,\boldsymbol{X})|S=0]P(S=0) \\ &= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &\quad + E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,\boldsymbol{X})P(A=a|S=0,\boldsymbol{X}) \\ &\quad + \{E(Y^a|S=0,A=a,\boldsymbol{X}) - b'(a,x)\}P(A\neq a|S=0,\boldsymbol{X})|S=0]P(S=0) \\ &= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &\quad + E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,\boldsymbol{X}) - b'(a,x)P(A\neq a|S=0,\boldsymbol{X})|S=0]P(S=0) \\ &= E_{\boldsymbol{X}}[E(Y|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &\quad + E_{\boldsymbol{X}}[E(Y|S=0,A=a,\boldsymbol{X}) - b'(a,x)P(A\neq a|S=0,\boldsymbol{X})|S=0]P(S=0) \end{aligned}$$

Identity (B.1) can be used as the basis for sensitivity analysis, substituting different plausible bias relationships for $b'(a,\boldsymbol{x})$, as was done by Brumback *et al.* (2004). In many

settings, it is unlikely that $b'(a,\boldsymbol{x})$ would have different signs for different $a$ (this would imply that within the same level of $\boldsymbol{X}$, individuals would have the largest outcome on the treatment they ended up on compared to other treatments). Among the various possible functional forms for the bias term presented in Brumback *et al.* (2004), we could assume bias would depend on measured covariates and take the form $b'(a,\boldsymbol{x}) = \boldsymbol{\beta}_{\boldsymbol{a}}\boldsymbol{X}$. Note the similarity to the 2-stage CCDS approach where a slightly different formulation of the bias term is estimated from the overlap region.

Identity (B.1) highlights that the bias from the naive "obs/rand" estimator in Section 2.4.6 that averages across randomized and observational estimates for randomized and observational units respectively is therefore:

$$
\begin{aligned}
&E_{\boldsymbol{X}}\Big[b'(a,x)P(A \neq a|S=0,\boldsymbol{X})\Big|S=0\Big]P(S=0) \\
&\quad = E_{\boldsymbol{X}}\Big[\Big\{E(Y^a|S=0,A=a,\boldsymbol{X}=\boldsymbol{x}) - E(Y^a|S=0,A\neq a,\boldsymbol{X}=\boldsymbol{x})\Big\} \\
&\qquad \times P(A \neq a|S=0,\boldsymbol{X})\Big|S=0\Big]P(S=0)
\end{aligned}
$$

## B.3 Proof for identification of $\psi_{\text{CCDS}}(a)$

$$E(Y^a) = E(Y^a|S=1)P(S=1) + E(Y^a|S=0)P(S=0) \tag{B.2}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,\boldsymbol{X})|S=1]P(S=1) + E_{\boldsymbol{X}}[E(Y^a|S=0,\boldsymbol{X})|S=0]P(S=0) \tag{B.3}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,\boldsymbol{X})|S=1]P(S=1) + E_{\boldsymbol{X}}[E(Y^a|S=1,\boldsymbol{X})|S=0]P(S=0) \tag{B.4}$$

$$\begin{aligned} &= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) \\ &\quad + E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=0]P(S=0) \end{aligned} \tag{B.5}$$

$$\begin{aligned} &= E_{\boldsymbol{X}|S=1}[E(Y^a|S=1,A=a,\boldsymbol{X}|S=1)]P(S=1) + E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,\boldsymbol{X}) \\ &\quad - \{E(Y^a|S=0,A=a,\boldsymbol{X}) - E(Y^a|S=1,A=a,\boldsymbol{X})\}|S=0]P(S=0) \end{aligned} \tag{B.6}$$

$$\begin{aligned} &= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|S=1]P(S=1) + E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,\boldsymbol{X}) \\ &\quad - \{E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\}|S=0]P(S=0) \end{aligned} \tag{B.7}$$

$$\begin{aligned} &= E_{\boldsymbol{X}}[\underbrace{E(Y|S=1,A=a,\boldsymbol{X})|S=1]}_{\text{(a) RCT contribution}}P(S=1) + E_{\boldsymbol{X}}[\underbrace{E(Y|S=0,A=a,\boldsymbol{X})}_{\substack{\text{(b) preliminary} \\ \text{observational contribution}}} \\ &\quad - \underbrace{\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\}}_{\text{(c) debiasing term for observational contribution}}|S=0]P(S=0) \end{aligned} \tag{B.8}$$

Lines (B.2) and (B.3) follow from the law of iterated expectations. Line (B.4) follows from Assumption 4 of conditional exchangeability for study selection; line (B.5) follows from the first part of Assumption 1b: $E(Y^a|S=1,A=a,\boldsymbol{X}) = E(Y^a|S=1,\boldsymbol{X})$; line (B.6) adds and subtracts the same term; line (B.7) then follows from the constant conditional bias part of Assumption 1b; line (B.8) follows from Assumptions 3 and 6 of SUTVA for treatment assignment and study selection; the final quantities are well-defined by the two positivity assumptions, 2 and 5.

One can alternatively identify treatment-specific means through different decompositions of the data in lines (B.3)-(B.4) (see Appendix B.4). Each functional implies different estimation strategies that rely on different auxiliary regression models.

## B.4 Estimators from alternative decompositions

One can identify PTSMs through functionals derived from alternative decompositions of the target population probability distribution. We present estimators $\hat{\psi}_2(a)$ and $\hat{\psi}_3(a)$ from two such alternative decompositions. Identification of $\psi_2(a)$ and $\psi_3(a)$ relies on a slightly different formulation of Assumption 1b: $b(a,x|R_{\text{obs-only}}=1) = b(a,x|R_{\text{overlap}}=1)$, i.e.,

$$\begin{aligned}
&E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) \\
&= E(Y^a|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X}) - E(Y^a|S=1,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})
\end{aligned}$$

Under this alternative formulation of Assumption 1b, along with Assumptions 2 - 5b, we can identify the causal estimand as follows:

- $$\begin{aligned}
\psi_2(a) =& E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,\boldsymbol{X})\Big|R_{\text{RCT}}=1\Big]P(R_{\text{RCT}}=1) \\
&+ E_{\boldsymbol{X}}\Big[E(Y|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X}) - \Big\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) \\
&\qquad - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\}\Big|R_{\text{obs-only}}=1\Big]P(R_{\text{obs-only}}=1)
\end{aligned}$$
- $$\begin{aligned}
\psi_3(a) =& E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,\boldsymbol{X})\Big|S=1\Big]P(S=1) \\
&+ E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big|S=0,R_{\text{overlap}}=1\Big]P(S=0,R_{\text{overlap}}=1) \\
&+ E_{\boldsymbol{X}}\Big[E(Y|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X}) - \Big\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) \\
&\qquad - E(Y|S=1,A=a,R_{\text{overlap}},\boldsymbol{X})=1\Big\}\Big|R_{\text{obs-only}}=1\Big]P(R_{\text{obs-only}}=1)
\end{aligned}$$

*Proof for $\psi_2(a)$:*

$$E(Y^a) = E(Y^a|R_{\text{RCT}}=1)P(R_{\text{RCT}}=1) + E(Y^a|R_{\text{obs-only}}=1)P(R_{\text{obs-only}}=1) \tag{B.9}$$

$$= E_{\boldsymbol{X}}[E(Y^a|R_{\text{RCT}}=1,\boldsymbol{X})|R_{\text{RCT}}=1]P(R_{\text{RCT}}=1)$$
$$+ E_{\boldsymbol{X}}[E(Y^a|R_{\text{obs-only}}=1,\boldsymbol{X})|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) \tag{B.10}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,R_{\text{RCT}}=1,\boldsymbol{X})|R_{\text{RCT}}=1]P(R_{\text{RCT}}=1)$$
$$+ E_{\boldsymbol{X}}[E(Y^a|S=1,R_{\text{obs-only}}=1,\boldsymbol{X})|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) \tag{B.11}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|R_{\text{RCT}}=1]P(R_{\text{RCT}}=1)$$
$$+ E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) \tag{B.12}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|R_{\text{RCT}}=1]P(R_{\text{RCT}}=1)$$
$$+ E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})$$
$$- \{E(Y^a|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})$$
$$- E(Y^a|S=1,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})\}|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) \tag{B.13}$$

$$= E_{\boldsymbol{X}}[E(Y^a|S=1,A=a,\boldsymbol{X})|R_{\text{RCT}}=1]P(R_{\text{RCT}}=1)$$
$$+ E_{\boldsymbol{X}}[E(Y^a|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})$$
$$- \{E(Y^a|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})$$
$$- E(Y^a|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\}|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) \tag{B.14}$$

$$= \underbrace{E_{\boldsymbol{X}}[E(Y|S=1,A=a,\boldsymbol{X})|R_{\text{RCT}}=1]}_{\text{(a) RCT contribution and observational contribution in region of overlap}} P(R_{\text{RCT}}=1) + E_{\boldsymbol{X}}[\underbrace{E(Y|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})}_{\text{(b) preliminary observational contribution in region of no overlap}}$$
$$\underbrace{-\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\}}_{\text{(c) debiasing term for observational contribution in region of no overlap}}|R_{\text{obs-only}}=1] \times P(R_{\text{obs-only}}=1) \tag{B.15}$$

*Proof for $\psi_3(a)$:*

$$
\begin{aligned}
E(Y^a) &= E(Y^a|S=1)P(S=1) + E(Y^a|S=0,R_{\text{overlap}}=1)P(S=0,R_{\text{overlap}}=1) \\
&\quad + E(Y^a|R_{\text{obs-only}}=1)P(R_{\text{obs-only}}=1) && \text{(B.16)} \\
&= E_{\boldsymbol{X}}[E(Y^a|S=1,\boldsymbol{X})|S=1]P(S=1) \\
&\quad + E_{\boldsymbol{X}}[E(Y^a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})|S=0,R_{\text{overlap}}=1]P(S=0,R_{\text{overlap}}=1) \\
&\quad + E_{\boldsymbol{X}}[E(Y^a|R_{\text{obs-only}}=1,\boldsymbol{X})|R_{\text{obs-only}}=1]P(R_{\text{obs-only}}=1) && \text{(B.17)} \\
&= \underbrace{E_{\boldsymbol{X}}[E(Y|S=1,A=a,\boldsymbol{X})|S=1]}_{\text{(a) RCT contribution}}P(S=1) \\
&\quad + \underbrace{E_{\boldsymbol{X}}[E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})|S=0,R_{\text{overlap}}=1]}_{\text{(a) observational contribution in region of overlap}}P(S=0,R_{\text{overlap}}=1) \\
&\quad + E_{\boldsymbol{X}}[\underbrace{E(Y|S=0,A=a,R_{\text{obs-only}}=1,\boldsymbol{X})}_{\text{(b) preliminary observational contribution in region of no overlap}} \\
&\qquad \underbrace{-\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\}|R_{\text{obs-only}}=1]}_{\text{(c) debiasing term for observational contribution in region of no overlap}} \times P(R_{\text{obs-only}}=1) \\
& && \text{(B.18)}
\end{aligned}
$$

As in the proof for $\psi_{\text{CCDS}}(a)$, lines (B.9), (B.10), and (B.16) follow from the law of iterated expectations. Line (B.11) follows from Assumption 4 of conditional exchangeability for study selection; line (B.12) follows from the first part of Assumption 1b: $E(Y^a|S=1,A=a,\boldsymbol{X}) = E(Y^a|S=1,\boldsymbol{X})$ (and the redundancy of $S=1$ and $\mathcal{R}_{\text{RCT}}$); line (B.13) adds and subtracts the same term; line (B.14) then follows from the constant conditional bias part of Assumption 1b; line (B.15) follows from Assumptions 3 and 6 of SUTVA for treatment assignment and study selection; the final quantities are well-defined by the two positivity assumptions, 2 and 5. For lines (B.17)-(B.18), the same steps seen in the proof of $\psi_2(a)$ were repeated to arrive at the final functional.

Each of these three functionals ($\psi_{\text{CCDS}}, \psi_2, \psi_3$) suggests slightly different estimation procedures that rely on different auxiliary regression models for different subsets of data. For example, the outcome regression estimators of $\psi_2$ and $\psi_3$ would be as follows:

$$\hat{\psi}_{2-OR}(a) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\hat{Q}_i(S_i=1, A_i=a, \boldsymbol{X}_i)\mathbb{1}(R_{\text{RCT}}=1)}_{\substack{\text{(a) RCT estimate and observational} \\ \text{estimate in region of overlap}}} + \underbrace{\hat{Q}_i(S_i=0, A_i=a, R_{\text{obs-only}}=1, \boldsymbol{X}_i)\mathbb{1}(R_{\text{i, obs-only}}=1)}_{\substack{\text{(b) preliminary observational estimate in} \\ \text{region of no overlap}}}$$
$$\underbrace{-\left\{\hat{Q}_i(S_i=0, A_i=a, R_{\text{overlap}}=1, \boldsymbol{X}_i) - \hat{Q}_i(S_i=1, A_i=a, R_{\text{overlap}}=1, \boldsymbol{X}_i)\right\}\mathbb{1}(R_{\text{i, obs-only}}=1)}_{\text{(c) debiasing term for observational estimate in region of no overlap}}$$

$$\hat{\psi}_{3-OR}(a) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\hat{Q}_i(S_i=1, A_i=a, \boldsymbol{X}_i)\mathbb{1}(S_i=1)}_{\text{(a) RCT estimate}} + \underbrace{\hat{Q}_i(S_i=1, A_i=a, R_{\text{overlap}}=1, \boldsymbol{X}_i)\mathbb{1}(S_i=0, R_{\text{i, overlap}}=1)}_{\substack{\text{(b1) preliminary observational estimate in} \\ \text{region of overlap}}}$$
$$+ \underbrace{\hat{Q}_i(S_i=0, A_i=a, R_{\text{obs-only}}=1, \boldsymbol{X}_i)\mathbb{1}(R_{\text{i, obs-only}}=1)}_{\substack{\text{(b2) preliminary observational estimate in} \\ \text{region of no overlap}}}$$
$$\underbrace{-\left\{\hat{Q}_i(S_i=0, A_i=a, R_{\text{overlap}}=1, \boldsymbol{X}_i) - \hat{Q}_i(S_i=1, A_i=a, R_{\text{overlap}}=1, \boldsymbol{X}_i)\right\}\mathbb{1}(R_{\text{i, obs-only}}=1)}_{\text{(c) debiasing term for observational estimate in region of no overlap}}$$

Choices between the two estimators here and the one presented in the main paper should rely on such considerations as efficiency and which regression models can be better-fit with the data (e.g., both $\hat{\psi}_{2-OR}(a)$ and $\hat{\psi}_{3-OR}(a)$ rely on preliminary observational estimates estimated from regressions fit to small subsets of the data). As a reminder, $\psi_{\text{CCDS}}(a)$ suggests an estimation procedure in which models are fit using: (a) all the randomized data to estimate potential outcomes in the randomized data, (b) all the observational data to estimate preliminary potential outcomes in the observational data, and (c) the randomized data in the overlap region and the observational data in the overlap region to estimate the debiasing term for preliminary observational data estimates.

In contrast, $\psi_2(a)$ suggests an estimation procedure in which regressions are fit using: (a) all the randomized data to estimate potential outcomes in the randomized and in the overlap region of the observational study, (b) the observational data in the nonoverlap region to estimate preliminary potential outcomes in the nonoverlap region of the observational study, and (c) the randomized data in the overlap region and the observational data in the overlap region to estimate the debiasing term.

Correspondingly, $\psi_3(a)$ suggests an estimation procedure in which regressions are fit using: (a) all the randomized data to estimate potential outcomes in the randomized population,

(b1) the randomized data in the overlap region to estimate potential outcomes in the overlap region of the observational study, (b2) the observational data in the nonoverlap region to estimate preliminary potential outcomes in the nonoverlap region of the observational study, and (c) the randomized data in the overlap region and the observational data in the overlap region to estimate the debiasing term.

The three estimators differ in the flexibility of their model specifications: the latter estimators let the covariate-outcome relationship differ in the overlap vs. nonoverlap regions. However, this flexibility comes at the cost of less information borrowing across the entire covariate distribution.

## B.5 Implementation

### B.5.1 CCDS-OR

Each of the outcome regressions in $\hat{\psi}_{\text{CCDS-OR}}(a)$ must appropriately capture treatment effect heterogeneity such as through including all relevant interaction terms in a least-squares regression model or by using flexible nonparametric approaches that discover effect heterogeneity in a data-driven fashion, such as machine learning algorithms (keeping in mind that many such approaches do not have convergence rates that result in $\sqrt{n}$-consistency). When fitting more complex models such as machine learning algorithms for the outcome regressions, there is a potential for overfitting to the trends in the overlap region when estimating the debiasing term (third term in $\hat{\psi}_{\text{CCDS-OR}}(a)$), even with regularization and cross-validation.

The CCDS framework can also be used to estimate conditional PTSMs for the CCDS-OR and 2-stage CCDS-OR estimators via a weighted average of randomized and debiased observational conditional means, weighted by the relative proportion of randomized and observational individuals in the target population. The CCDS-OR conditional PTSM estimator

is as follows:

$$\hat{\psi}_{\text{CCDS-OR}}(a,\boldsymbol{x}) = \underbrace{\frac{n_{\text{rand}}}{n}\hat{Q}(S=1,A=a,\boldsymbol{X}=\boldsymbol{x})}_{\text{(a) RCT estimate}} + \underbrace{\frac{n_{\text{obs}}}{n}\hat{Q}(S=0,A=a,\boldsymbol{X}=\boldsymbol{x})}_{\text{(b) preliminary observational estimate}}$$
$$\underbrace{-\frac{n_{\text{obs}}}{n}\Big\{\hat{Q}(S=0,A=a,\hat{R}_{\text{overlap}}=1,\boldsymbol{X}=\boldsymbol{x}) - \hat{Q}(S=1,A=a,\hat{R}_{\text{overlap}}=1,\boldsymbol{X}=\boldsymbol{x})\Big\}}_{\text{(c) debiasing term for observational estimate}}$$

Rather than simply using $n_{\text{study}}/n$, these weights can also be replaced by sampling weights.

### B.5.2 2-stage CCDS

In the second stage of the 2-stage CCDS estimator, a simple $\hat{g}()$ function such as $\hat{g}(\boldsymbol{X}) = \boldsymbol{X}^T\hat{\theta}$ can prevent overfitting to the overlap region and thus provide added stability to estimating bias, particularly when fitting more complex $\hat{Q}(S,A,R,\boldsymbol{X})$ regressions in the first stage. Substantive knowledge can also inform choice of the $\hat{g}()$ function, such as knowledge of which measured covariates can serve as a proxy for unmeasured confounders.

In studies with fewer treatment groups and thus more data in each one, it may be beneficial to subset to randomized overlap region data in a given treatment group for bias estimation to make sure to fully capture treatment effect heterogeneity (though this approach precludes borrowing strength across treatment groups). Step 2 of the 2-stage CCDS estimator then becomes:

(2) $\hat{b}'(S_i=1,a,\boldsymbol{X}_i) = \frac{\hat{w}_{\text{bias}}(S_i,A_i,\boldsymbol{X}_i)}{\sum_{i=1}^{n}\hat{w}_{\text{bias}}(S_i,A_i,\boldsymbol{X}_i)}\hat{g}(\boldsymbol{X}_i)$ with

$$\hat{w}_{\text{bias}}(S_i,A_i,\boldsymbol{X}_i) = \frac{\mathbb{1}(S_i=1,A_i=a,R_{\text{overlap, i}}=1)\hat{P}(S_i=0|\boldsymbol{X}_i)}{\hat{P}(R_{\text{overlap, i}}=1|S_i=1,\boldsymbol{X}_i)\hat{P}(S_i=1|\boldsymbol{X}_i)\hat{P}(A_i=a|S_i=1,R_{\text{overlap, i}}=1,\boldsymbol{X}_i)}$$

### B.5.3 CCDS-IPW

To circumvent unstable weights for CCDS-IPW and other novel estimators using weights, propensity scores and their products used in weight denominators can be trimmed. However, trimming weights effectively changes the estimand of interest, thus requiring a bias-variance tradeoff.

## B.6 2-stage whole data outcome regression estimator

### B.6.1 Estimator

An alternative to the constant conditional bias assumption is to instead extrapolate from the randomized study to regions not supported in the randomized study covariate distribution, $\mathcal{R}_{\text{obs-only}}$. If we believe that we can reliably extrapolate from the randomized data for the purpose of debiasing term estimation (although we are not confident enough to directly extrapolate potential outcomes), the 2-stage whole data (WD) outcome regression estimator would provide more power than the 2-stage CCDS approach by not restricting debiasing term estimation to the overlap region:

(1) $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i) = \hat{Q}_i(S = 0, A = a, \boldsymbol{X})\mathbb{1}(S_i = 1) - \hat{Q}_i(S = 1, A = a, \boldsymbol{X})\mathbb{1}(S_i = 1)$

(2) $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i) = \frac{\hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i)} \hat{g}(\boldsymbol{X}_i)$ with $\hat{w}_{\text{bias}}(S_i, \boldsymbol{X}_i) = \frac{\mathbb{1}(S_i=1)\hat{P}(S_i=0|\boldsymbol{X}_i)}{\hat{P}(S_i=1|\boldsymbol{X}_i)}$.

One could likewise subset to randomized data in a given treatment group for bias estimation. The 2-stage WD estimator then becomes:

(1) $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i) = \hat{Q}_i(S = 0, A = a, \boldsymbol{X})\mathbb{1}(S_i = 1, A_i = a) - \hat{Q}_i(S = 1, A = a, \boldsymbol{X})\mathbb{1}(S_i = 1, A_i = a)$

(2) $\hat{b}'(S_i = 1, a, \boldsymbol{X}_i) = \frac{\hat{w}_{\text{bias}}(S_i, A_i, \boldsymbol{X}_i)}{\sum_{i=1}^{n} \hat{w}_{\text{bias}}(S_i, A_i, \boldsymbol{X}_i)} \hat{g}(\boldsymbol{X}_i)$ with $\hat{w}_{\text{bias}}(S_i, A_i, \boldsymbol{X}_i) = \frac{\mathbb{1}(S_i=1, A_i=a)\hat{P}(S_i=0|\boldsymbol{X}_i)}{\hat{P}(S_i=1|\boldsymbol{X}_i)\hat{P}(A_i=a|S_i=1, A_i=a, \boldsymbol{X}_i)}$.

A similar approach was taken by Kallus *et al.* (2018) for estimating target population conditional average treatment effects for a target population represented by the observational data, using $Y_i\mathbb{1}(S_i = 1, A_i = a)/\hat{P}(A_i = a|S_i = 1, \boldsymbol{X}_i)$ instead of $\hat{Q}_i(S = 1, A = a, \boldsymbol{X})\mathbb{1}(S_i = 1)$ in Stage (1) and not weighting Stage (2). Therefore, the Kallus et al. 2-stage approach optimizes for mean squared error across the covariate distribution in the randomized group rather than for that in the observational group, a covariate distribution that does not represent the one over which we want to minimize bias. However, it does not suffer from potentially increased variability due to the weights. The Kallus et al. 2-stage approach does not directly extend to estimating PTSMs.

### B.6.2 Simulation results

With correctly specified models, all novel estimators including the 2-stage WD approach were able to decrease unmeasured confounding bias and the 2-stage WD approach was the most efficient novel outcome regression estimator because it used more data to fit regressions compared to CCDS estimators (Appendix Figure B.1). However, when (incorrectly) fitting main terms regressions, just as with the rand estimator, extrapolation became an issue for the 2-stage WD estimator. In fact, with linear additive models like the correctly specified and main terms regressions, the 2-stage WD estimator is numerically equivalent to the rand estimator due to the linearity and additivity of the models. Even with ensemble approaches, 2-stage WD's bias tended to be similar to that of the rand estimator. Because of the 2-stage WD estimator's sensitivity to model misspecificiation, we do not generally recommend using this estimator. The estimator's poor performance highlights the importance of focusing on the overlap region for estimating unmeasured confounding bias.

## B.7 Proof for $\hat{\psi}_{\text{CCDS-IPW}}(a)$

$\psi_{\text{CCDS}}(a)$ consists of four components:

$$\begin{aligned}\psi_{\text{CCDS}}(a) = &\underbrace{E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,\boldsymbol{X})\Big|S=1\Big]}_{(1)}P(S=1)\\ &+\underbrace{E_{\boldsymbol{X}}\Big[E(Y|S=0,A=a,\boldsymbol{X})|S=0\Big]}_{(2)}P(S=0)\\ &-\Big\{\underbrace{E_{\boldsymbol{X}}\Big[E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})|S=0\Big]}_{(3)}\\ &\quad-\underbrace{E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big|S=0\Big]}_{(4)}\Big\}P(S=0)\end{aligned}$$

We can then identify each of the conditional distributions via the following propensity decomposition (using conditional probability laws and positivity assumptions). For example,

for component (4):

$$(4) = E_{\boldsymbol{X}}\Big[E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big|S=0\Big] \tag{B.19}$$

$$= \frac{1}{P(S=0)}E_{\boldsymbol{X}}\Big[\mathbb{1}(S=0)E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big] \tag{B.20}$$

$$= \frac{1}{P(S=0)}E_{\boldsymbol{X}}\left[P(S=0|\boldsymbol{X})\frac{E(Y\mathbb{1}(S=1,A=a,R_{\text{overlap}}=1)|\boldsymbol{X})}{P(S=1,R_{\text{overlap}}=1|\boldsymbol{X})P(A=a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})}\right] \tag{B.21}$$

$$= \frac{1}{P(S=0)}E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=1,A=a,R_{\text{overlap}}=1)P(S=0|\boldsymbol{X})}{P(S=1,R_{\text{overlap}}=1|\boldsymbol{X})P(A=a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})}\right)\right] \tag{B.22}$$

For weight stabilization, we can also replace $\frac{1}{P(S=0)}$ with

$$E(w_4)^{-1} = \left[E_{\boldsymbol{X}}\left(\frac{\mathbb{1}(S=1,A=a,R_{\text{overlap}}=1)P(S=0|\boldsymbol{X})}{P(S=1,R_{\text{overlap}}=1|\boldsymbol{X})P(A=a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})}\right)\right]^{-1}$$

since

$$E(w_4) = E_{\boldsymbol{X}}\left(\frac{E(\mathbb{1}(S=1,A=a,R_{\text{overlap}}=1)|\boldsymbol{X})P(S=0|\boldsymbol{X})}{P(S=1,R_{\text{overlap}}=1|\boldsymbol{X})P(A=a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})}\right)$$

$$= E_{\boldsymbol{X}}\Big(P(S=0|\boldsymbol{X})\Big) = P(S=0)$$

This weight stabilization creates more stability for estimation and ensure estimates are in the support of the outcome variable. Note: $E(w_4)$ can cancel with $E(P(S=0|\boldsymbol{X}))$ in line (B.22) but doing so would remove the weight stabilization.

We can similarly identify each of the conditional distributions in (1) - (3) through the following propensity score decompositions:

$$(1) = E(w_1)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=1,A=a)P(S=1|\boldsymbol{X})}{P(S=1|\boldsymbol{X})P(A=a|S=1,\boldsymbol{X})}\right)\right] \tag{B.23}$$

$$= E(w_1)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=1,A=a)}{P(A=a|S=1,\boldsymbol{X})}\right)\right] \tag{B.24}$$

$$(2) = E(w_2)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=0,A=a)P(S=0|\boldsymbol{X})}{P(S=0|\boldsymbol{X})P(A=a|S=0,\boldsymbol{X})}\right)\right] \tag{B.25}$$

$$= E(w_2)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=0,A=a)}{P(A=a|S=0,\boldsymbol{X})}\right)\right] \tag{B.26}$$

$$(3) = E(w_3)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=0,A=a,R_{\text{overlap}}=1)P(S=0|\boldsymbol{X})}{P(S=0,R_{\text{overlap}}=1|\boldsymbol{X})P(A=a|S=0,R_{\text{overlap}}=1,\boldsymbol{X})}\right)\right] \tag{B.27}$$

$$= E(w_3)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=0,A=a,R_{\text{overlap}}=1)P(S=0|\boldsymbol{X})}{P(S=0|\boldsymbol{X})P(R_{\text{overlap}}=1|S=0,\boldsymbol{X})P(A=a|S=0,R_{\text{overlap}}=1,\boldsymbol{X})}\right)\right] \tag{B.28}$$

$$= E(w_3)^{-1} E_{\boldsymbol{X}}\left[E\left(\frac{Y\mathbb{1}(S=0,A=a,R_{\text{overlap}}=1)}{P(R_{\text{overlap}}=1|S=0,\boldsymbol{X})P(A=a|S=0,R_{\text{overlap}}=1,\boldsymbol{X})}\right)\right] \tag{B.29}$$

$$\tag{B.30}$$

where

$$w_1 = \frac{\mathbb{1}(S=1,A=a)}{P(A=a|S=1,\boldsymbol{X})}$$

$$w_2 = \frac{\mathbb{1}(S=0,A=a)}{P(A=a|S=0,\boldsymbol{X})}$$

$$w_3 = \frac{\mathbb{1}(S=0,A=a,R_{\text{overlap}}=1)}{P(R_{\text{overlap}}=1|S=0,\boldsymbol{X})P(A=a|S=0,R_{\text{overlap}}=1,\boldsymbol{X})}$$

$$w_4 = \frac{\mathbb{1}(S=1,A=a,R_{\text{overlap}}=1)[1-P(S=1|\boldsymbol{X})]}{P(S=1|\boldsymbol{X})P(R_{\text{overlap}}=1|S=1,\boldsymbol{X})P(A=a|S=1,R_{\text{overlap}}=1,\boldsymbol{X})}$$

## B.8 CCDS influence function

To derive the influence function for $\psi_{\text{CCDS}}(a)$, we first derive the influence function for each of its four conditional means:

(1) For $\chi_1(a) = E_{\boldsymbol{X}}[E(Y|S=1,A=a,\boldsymbol{X})|S=1]$,

$$\chi_1'(a) = \frac{1}{P(S=1)}\Big[w_1\Big\{Y - E(Y|S=1,A=a,\boldsymbol{X})\Big\} + S\Big\{E(Y|S=1,A=a,\boldsymbol{X}) - \chi_1(a)\Big\}\Big]$$

(2) For $\chi_2(a) = E_{\boldsymbol{X}}[E(Y|S=0,A=a,\boldsymbol{X})|S=0]$,

$$\chi_2'(a) = \frac{1}{P(S=0)}\Big[w_2\Big\{Y - E(Y|S=0,A=a,\boldsymbol{X})\Big\} + (1-S)\Big\{E(Y|S=0,A=a,\boldsymbol{X}) - \chi_2(a)\Big\}\Big]$$

(3) For $\chi_3(a) = E_{\boldsymbol{X}}[E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})|S=0]$,

$$\begin{aligned}\chi_3'(a) = \frac{1}{P(S=0)}\Big[&w_3\Big\{Y - E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\} \\ &+(1-S)\Big\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - \chi_3(a)\Big\}\Big]\end{aligned}$$

(4) For $\chi_4(a) = E_{\boldsymbol{X}}[E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})|S=0]$,

$$\begin{aligned}\chi_4'(a) = \frac{1}{P(S=0)}\Big[&w_4\Big\{Y - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\} \\ &+(1-S)\Big\{E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - \chi_4(a)\Big\}\Big]\end{aligned}$$

where probabilities and expectations are taken under the true model and weights are as previously defined.

The joint influence function will then be the reweighted (by $P(S=1)$ or $P(S=0)$) sum of the 4 conditional mean influence functions:

$$\begin{aligned}\chi'(a) = &w_1\Big\{Y - E(Y|S=1,A=a,\boldsymbol{X})\Big\} + S\Big\{E(Y|S=1,A=a,\boldsymbol{X}) - \chi_1(a)\Big\} \\ &+ w_2\Big\{Y - E(Y|S=0,A=a,\boldsymbol{X})\Big\} + (1-S)\Big\{E(Y|S=0,A=a,\boldsymbol{X}) - \chi_2(a)\Big\} \\ &- w_3\Big\{Y - E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\} \\ &\quad - (1-S)\Big\{E(Y|S=0,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - \chi_3(a)\Big\} \\ &+ w_4\Big\{Y - E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X})\Big\} \\ &\quad + (1-S)\Big\{E(Y|S=1,A=a,R_{\text{overlap}}=1,\boldsymbol{X}) - \chi_4(a)\Big\}\end{aligned}$$

## B.9 Supplemental simulation descriptions and results

### B.9.1 Further implementation details

Ensemble regressions were implemented using the `SuperLearner` package (Polley *et al.*, 2019) and consisted of `SL.glm`, `SL.glm.interact`, `SL.glmnet` with $\alpha = 0.5$, `SL.ranger` with 300 trees and a minimum node size of 5% of the sample being fit, `SL.nnet` with 2 hidden layers, `SL.earth`, `SL.gam`, and `SL.kernelKnn`. For primary results, we conservatively estimated the overlap region using $\alpha = 1\% \times range(logit(\pi_S))$ and $\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}})$, i.e., at least 1% of observations in a given treatment group must fall within 1% intervals of the logit propensity score.

### B.9.2 Further descriptions of the data generating mechanism

The data generating mechanism (DGM) resulted in positivity of selection violation (Figure B.2B), the confounders having varying strengths of confounding, there being relatively strong unmeasured confounding ($U$ had the second largest impact on treatment and outcome values), the conditional outcome relationship in $\mathcal{R}_{\text{overlap}}$ in the randomized data not fully extrapolating well to $\mathcal{R}_{\text{obs-only}}$ unless precisely the correct outcome model was fit, and observed covariates ($X_1$) differing in distribution across randomized and observational data. As a result, randomized and observational data each displayed external validity bias for estimating PTSM and PATE, and observational data likewise displayed internal validity bias due to measured and unmeasured confounding (Appendix Table B.1). With these specifications, there likewise was discrepancy between true randomized and observational study population treatment-specific means (STSMs) and study population average treatment effects (SATEs) (Table B.1).

This DGM also ensured that identifiability assumptions held, namely:

1. The randomized group had no unmeasured confounding and the distribution of $U$ was the same in $\mathcal{R}_{\text{overlap}}$ as $\mathcal{R}_{\text{obs-only}}$ conditioning on measured covariates; thus, the constant conditional bias assumption was satisfied (bias was in fact constant, not just conditionally constant; it was equal to $E(10U)$).

2. All study/treatment groups had a positive probability of receiving each treatment (treatment propensities are bounded between approximately 0.2-0.8, Appendix Figure B.2A).

3. Observations were independent.

4. The unmeasured covariate did not confound the relationship between outcome and study selection.

5. $\mathcal{R}_{\text{overlap}}$ was not a null set, Appendix Figure B.2.

6. The same outcome model held for both randomized and observational data.

We likewise investigated the impacts of violations to these assumptions and of other DGM and regression specifications on CCDS estimators' performance. Working off the base case (a true outcome model that contains higher order terms, unmeasured confounding, and positivity of selection violation; regressions fit with either linear models or an ensemble approach using either the true overlap region or estimating the overlap region using $\alpha = 1\% \times range(logit(\pi_S))$ and $\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}})$), we assessed the following settings: 6 model fit specifications (main effects, squared terms, correctly specified, ksvm, and two ensembles), 5 target sample sizes ($n = 200, 2000, 10000, 20000, 50000$), a constant bias violation setting, 4 unmeasured confounding settings, 3 overlap settings, 3 ratios of $n_{\text{RCT}}$ to $n_{\text{obs}}$, 3 positivity of selection violation settings, 2 exchangeability of study selection violations, 5 overlap region determination settings, and 3 propensity for selection relationships. We also examined 6 alternative outcome models (main effects model, more complex effect heterogeneity, knot, more severe knot, knot inside overlap region, $U \times X_1$ interaction).

### B.9.3 Overlap region specifications

We examined a range of overlap region specifications (Appendix Table B.2), with $\alpha$ and $\beta$ overlap region hyperparameters set on the propensity and log propensity scales. The overlap region specifications used in the base case ($\alpha = 1\% \times range(logit(\pi_S))$ and $\beta = 1\% \times$

| | Population | RCT | | Obs | |
|---|---|---|---|---|---|
| | Truth | Truth | Observed | Truth | Observed |
| $E(Y^1)$ | 5.09 | 13.49 | 13.48 | 3.00 | 1.47 |
| $E(Y^2)$ | 2.09 | 15.11 | 15.11 | -1.16 | 1.65 |
| $E(Y^1 - Y^2)$ | 3.00 | -1.63 | -1.62 | 4.16 | -0.19 |

**Table (B.1)** *Population and sample true potential outcome means and means observed in each treatment group*

| | Mean overlap | | |
|---|---|---|---|
| | Obs | RCT | Total |
| Truth | 38% | 50% | 40% |
| $\alpha = 1\% \times range(\pi_S) = 0.01$<br>$\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}})) = 20$ | 24% | 29% | 25% |
| $\alpha = 1\% \times range(logit(\pi_S)) = 0.3$<br>$\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}})) = 20$ | 35% | 48% | 38% |
| $\alpha = 2\% \times range(\pi_S) = 0.02$<br>$\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}}) = 20$ | 38% | 42% | 39% |
| $\alpha = 2\% \times range(logit(\pi_S)) = 0.6$<br>$\beta = 1\% \times min(n_{\text{obs}}, n_{\text{rand}}) = 20$ | 50% | 61% | 52% |
| $\alpha = 10\% \times range(logit(\pi_S)) = 3$<br>$\beta = 4\% \times min(n_{\text{obs}}, n_{\text{rand}}) = 78$ | 91% | 89% | 91% |

**Table (B.2)** *Overlap region specifications*

$min(n_{\text{obs}}, n_{\text{rand}}))$ were the closest to approximating the true overlap region, particularly for randomized data. Across the range of scenarios examined, which spanned underestimating to grossly overestimating the overlap region, surprisingly, bias and RMSE were minimally impacted (Appendix Figure B.3).

### B.9.4 Different degrees of overlap (positivity of study selection violation)

We changed the size of $\mathcal{R}_{\text{overlap}}$ from the default of $QNorm(0.5) = 0 \leq X_1 \leq 1.28 = QNorm(0.9)$ to $QNorm(0.7) = 0.52 \leq X_1 \leq 1.28 = QNorm(0.9)$ for less overlap and $QNorm(0.1) = -1.28 \leq X_1 \leq 1.28 = QNorm(0.9)$ for more overlap. These changes resulted in different proportions of observational data falling in the overlap region (13%, 38%, and 88%, respectively), but always retained 50% of randomized data in the overlap region.

With more overlap, all estimators besides obs/rand were able to shrink bias close to zero

(Appendix Figure B.4). Novel estimators had a larger region in which to estimate bias and the rand estimator was able to extrapolate better since more of the target population was in its region of support. With less overlap, all novel estimators' bias remained minimal though variance increased (most starkly for the CCDS-OR/CCDS-AIPW estimators with ensemble models); rand model bias increased sharply for estimating PTSMs. The DGM allowed PATEs to be extrapolated from the randomized data, so a corresponding bias increase was not observed for PATEs. The CCDS-IPW's RMSE remained the least affected among all novel estimators.

### B.9.5 Different ratios of $n_{\text{RCT}} : n_{\text{obs}}$

The base-case ratio of $n_{\text{RCT}} : n_{\text{obs}}$ was 1:4. We also examined 1:1 and 1:30 ratios. To maintain the same overlap region across all settings, we changed the overlap region bounds to $QNorm(0.18) = -0.92 \leq X_1 \leq 2.05 = QNorm(0.98)$ .

Bias and RMSE largely decreased across all estimators as the ratio of randomized to observational observations increased (Appendix Figure B.5). As the randomized observations comprised a larger portion of the target sample, the bias from using the rand and obs/rand estimators also correspondingly decreased. With ensemble models, the rate of bias and RMSE decrease for rand and novel estimators exceeded that of the obs/rand estimator, highlighting the large impact of having more randomized data when overlap is small. For correctly specified models, the rate of RMSE decrease exceeded that of the obs/rand models. Estimators' relative performance largely remained the same.

### B.9.6 Varying sample sizes

We examined sample sizes $n = 2,000$; $10,000$; and $50,000$. With correctly specified models, although all estimators' bias was lower than that of the obs/rand estimator, the CCDS-OR and CCDS-AIPW estimators retained a non-trivial amount of bias that minimally decreased with larger sample sizes due to remnant modeling bias from fitting complex models in the small overlap region (Appendix Figure B.6). With smaller sample sizes ($n = 2000$; $n_{\text{RCT}} = 500$), the

RMSE of all novel estimators and the rand estimator exceeded that of the obs/rand estimator. With ensemble models, all novel estimators' bias was below that of the obs/rand estimator. RMSE, however, only dropped below that of the obs/rand estimator with $n = 10,000$ for the PATE (except for the CCDS-IPW estimator, whose RMSE was lower even with $n = 2,000$).

### B.9.7 Varying strengths of unmeasured confounding

We examined four settings: no unmeasured confounding (in which $U$ was included as a measured covariate) and three levels of unmeasured confounding, for which the $U$ coefficient in the $P(Y|S,X,U)$ model was varied from 0.1 to 0.625 to 1.5 and the $U$ coefficient in the $P(A|S,X,U)$ model was varied from 5 to 10 to 20 for low confounding, default confounding, and high confounding, respectively. Results were similar for no and little unmeasured confounding. As unmeasured confounding bias increased, with correctly specified models, there was no corresponding increase in bias across novel estimators, though variance increased, reflecting more uncertainty in settings with more unmeasured confounding (Appendix Figure B.7). With ensemble models, there was a small increase in bias with more confounding. This increase was smaller for CCDS estimators than for the rand estimator.

### B.9.8 Constant conditional bias assumption violation

To violate the constant conditional bias assumption, the amount of unmeasured confounding bias was varied in a way that is not predictable from the trends observed in the overlap region (note that the overlap region lower bound is at $X_1 = 0$): $E(Y^1) = E_{\text{base}}(Y^1) - 45 * U * (X_1 + 0.5) * I(X_1 < -0.5)$ and $E(Y^2) = E_{\text{base}}(Y^2) - 30 * U * (X_1 + 0.5) * I(X_1 < -0.5)$ where $E_{\text{base}}(Y^a)$ corresponds to the base-case potential outcome model.

When the bias relationship observed in the overlap region differed from that outside the overlap region, bias and RMSE increased for each of the estimators corresponding to the amount of extra unmeasured confounding bias observed in the observational data which cannot be estimated from the overlap region (Appendix Figure B.8). The rand estimator was likewise not able to extrapolate well outside the overlap region even with a correctly

specified model. Novel estimators' bias remained below that of the obs/rand estimator as they removed the portion of unmeasured confounding bias that was estimable from the overlap region. This bias increase observed with constant conditional bias assumption violation thus accommodates the extra unmeasured confounding bias which cannot be removed, reflecting that these novel methods can only remove bias estimable from the overlap region: they rely on the constant conditional bias assumption. With ensemble models, surprisingly, the rand estimator's PATE bias and RMSE decreased with the constant bias violation, likely reflecting a function of the DGM as this result was not observed for PTSMs: when estimating PTSMs, the rand estimator was the most affected by the assumption violation.

### B.9.9 Exchangeability of study selection violation

We examined this assumption violation through two approaches. In the first, $P(S|\boldsymbol{X}, U)$ was changed to be a function of $U$ in the overlap region, $P(S = 1|U) = 0.125 + 0.25U$, and remained deterministically 0 or 1 outside the overlap region; $P(S = 1)$ remained at 0.20. When study selection was a function of an unmeasured outcome determinant, all estimators besides the obs/rand estimator showed an increase in bias for PTSMs (but not for PATEs due to $U$ not being an unmeasured effect modifier) such that bias from all estimators exceeded that of the obs/rand estimator (Appendix Figure B.9). Interestingly, for estimating $E(Y^1)$, the rand estimator's bias decreased slightly, though this was not observed for $E(Y^2)$ estimation.

In the second violation assessment, $U$ was a function of $X_1$, which determines overlap region membership: $U \sim Binom(p_U)$ where $p_U = expit(30X_1)$. Hence, $U$ was an unmeasured effect modifier with different distributions in the randomized vs. observational data. This resulted in the distribution of $U$ differing in the overlap vs. nonoverlap regions, so randomized estimates would not represent the truth in the overlap region. Hence, the CCDS estimators' bias increased with the assumption violation – when rand estimates are biased for the target population quantities in the overlap region, CCDS estimators are not be able to properly debias (Appendix Figure B.9). Likewise, the rand estimates' bias also increased in all but the ensemble estimating the PATE, in which surprisingly bias actually decreased,

likely due to bias cancellation between treatment groups.

### B.9.10 Alternative data generating mechanisms

We examined alternative DGMs for $Y$, $S$, and $A$ such as using simple main terms linear models (excluding $X_1^3$ terms), including more complex effect heterogeneity such that the PATE was not extrapolatable from the randomized data ($\mu_Y = -1.5 - 3A + 4X_1 + 4X_2 + 3X_3 + 2X_4 + 2(X_1+1)^3 + 4AX_1 + 2A(X_1+1)^3 + 10U$), using knot terms ($\mu_Y = \mu_{\text{Y, base}} - 15 * I(X_1 < -1) * (X_1+1) - 15 * I(X_1 < -1) * (X_1+1) * A$, $\mu_Y = \mu_{\text{Y, base}} - 45 * I(X_1 < -0.5) * (X_1+0.5) + 15 * I(X_1 < -0.5) * (X_1+0.5) * A$, $\mu_Y = \mu_{\text{Y, base}} - 2 * I(X_1 < 0.5) * (X_1 - 0.5) - 2 * I(X_1 < 0.5) * (X_1 - 0.5) * A$), and including an interaction between the unmeasured confounder and a measured covariate ($\mu_Y = \mu_{\text{Y, base}} + 2 * U * X_1 + 1 * U * X_1 * A$), where $\mu_{\text{Y, base}}$ is the outcome mean in the base-case. Conclusions remained similar across all examined DGMs.

## B.10 Supplemental Medicaid results

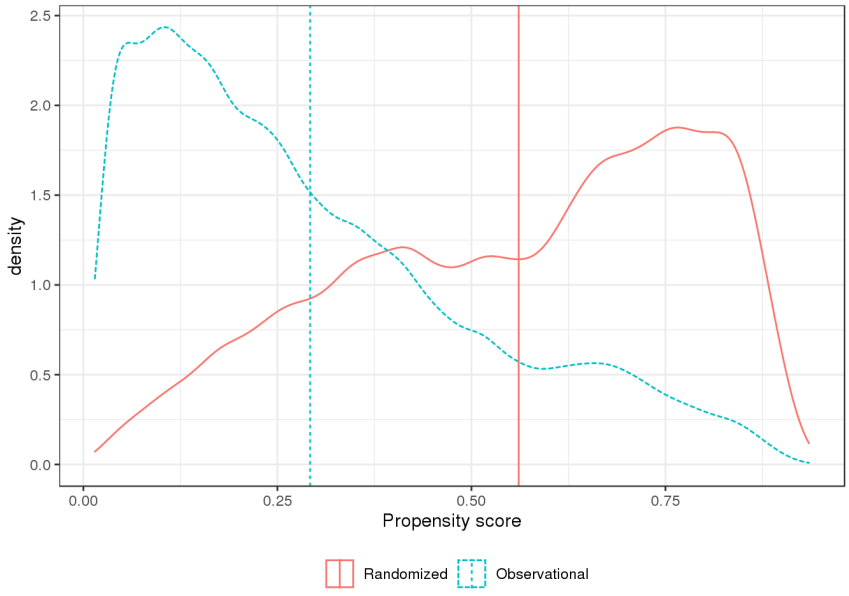



**Figure (B.10)** *Propensity for selection into the randomized group*

*The plot displays the density and mean, estimated using a linear regression.*

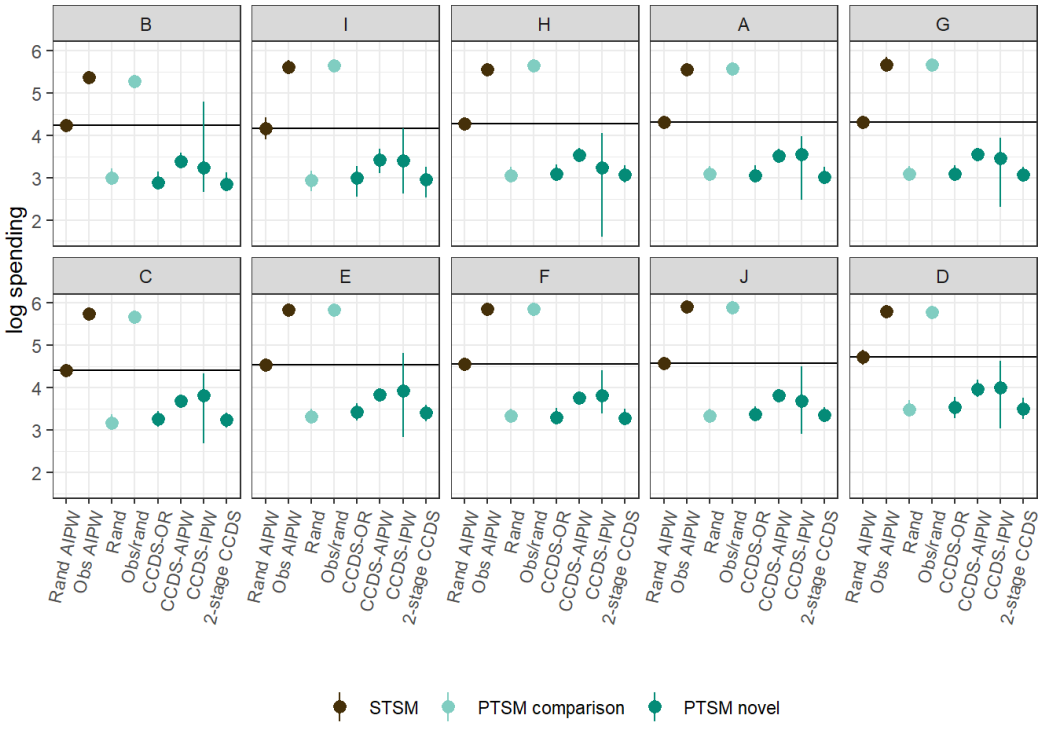



**Figure (B.11)** *STSMs and PTSMs across health plans for all estimators, with 95% confidence intervals multiplicity-adjusted with the Bonferroni correction*

## B.11 CCDS extensions

There are many possible future extensions to the CCDS estimation approach.

**Positivity of treatment assignment violations.** The novel estimators could be extended to similarly weaken Assumption 2, positivity of treatment assignment, although care must be taken to distinguish between structural and empirical/practical violations of positivity. The causal estimand would not be well-defined for individuals who could not receive a given intervention (if there was a structural positivity violation). Unless we were willing to change the causal estimand, we would need to still assume that structural positivity holds (which is an untestable population characteristic), but do not need to assume that empirical positivity holds (which is a finite sample characteristic).

**Multiple studies.** This approach can be extended to a collection of more than two studies with a common set of covariates ($\boldsymbol{X}$). Individual-level subject data could be separated into a

| Characteristic | Randomized | Observational | p-value |
|---|---|---|---|
| Sample size | 65591 | 98232 | |
| 6 month spending (mean (SD)) | 3052 (10089) | 2796 (6756) | <0.001 |
| Plan (n (%)) | | | <0.001 |
| A | 8510 ( 13.0) | 9879 ( 10.1) | |
| B | 7814 ( 11.9) | 6390 ( 6.5) | |
| C | 6195 ( 9.4) | 6200 ( 6.3) | |
| D | 2626 ( 4.0) | 18149 ( 18.5) | |
| E | 6770 ( 10.3) | 11302 ( 11.5) | |
| F | 8055 ( 12.3) | 17673 ( 18.0) | |
| G | 8439 ( 12.9) | 5689 ( 5.8) | |
| H | 7062 ( 10.8) | 6833 ( 7.0) | |
| I | 1420 ( 2.2) | 3402 ( 3.5) | |
| J | 8700 ( 13.3) | 12715 ( 12.9) | |
| Age (mean (SD)) | 35.55 (12.65) | 34.26 (12.75) | <0.001 |
| Female (n (%)) | 26370 ( 40.2) | 58076 ( 59.1) | <0.001 |
| Race (n (%)) | | | <0.001 |
| White non-Hispanic | 17808 ( 27.2) | 33258 ( 33.9) | |
| Black | 33853 ( 51.6) | 29347 ( 29.9) | |
| Asian or Pacific Islander | 3020 ( 4.6) | 19223 ( 19.6) | |
| American Indian or Alaskan Native | 1126 ( 1.7) | 1892 ( 1.9) | |
| Other | 9784 ( 14.9) | 14512 ( 14.8) | |
| County (n (%)) | | | <0.001 |
| Bronx | 16423 ( 25.0) | 21942 ( 22.3) | |
| Brooklyn | 21044 ( 32.1) | 32307 ( 32.9) | |
| Manhattan | 13281 ( 20.2) | 13002 ( 13.2) | |
| Queens | 12679 ( 19.3) | 27544 ( 28.0) | |
| Staten Island | 2164 ( 3.3) | 3437 ( 3.5) | |
| Aid group (n (%)) | | | <0.001 |
| MA SN adult | 31430 ( 47.9) | 56210 ( 57.2) | |
| MA SN child | 102 ( 0.2) | 339 ( 0.3) | |
| MA SSI blind | 714 ( 1.1) | 431 ( 0.4) | |
| MA TANF adult | 10867 ( 16.6) | 27553 ( 28.0) | |
| MA TANF child | 931 ( 1.4) | 2246 ( 2.3) | |
| SN adult | 15573 ( 23.7) | 6267 ( 6.4) | |
| SN child | 99 ( 0.2) | 125 ( 0.1) | |
| SSI blind | 5114 ( 7.8) | 1358 ( 1.4) | |
| TANF adult | 648 ( 1.0) | 3520 ( 3.6) | |
| TANF child | 65 ( 0.1) | 146 ( 0.1) | |
| Other | 48 ( 0.1) | 37 ( 0.0) | |
| Eligible for SSI (n (%)) | 5840 ( 8.9) | 1797 ( 1.8) | <0.001 |
| Baseline spending decile (mean (SD)) | 6.24 (3.31) | 3.88 (3.40) | <0.001 |
| Missing baseline spending (n (%)) | 839 ( 1.3) | 715 ( 0.7) | <0.001 |
| Percent neighborhood poverty (mean (SD)) | 0.24 (0.08) | 0.23 (0.08) | <0.001 |

*NOTE: The p-values correspond to a t-test for continuous variables and a chi-squared test for categorical variables, with a continuity correction.*

*Abbreviations: MA = Medicare Advantage; SD = standard deviation; SN = safety net; SSI = social security income; TANF = Temporary Assistance for Needy Families.*

**Table (B.3)** *Characteristics of randomized and observational Medicaid groups*

"randomized" group of individuals containing individuals from all studies which meet the mean conditional exchangeability requirement, and an "observational" group of individuals from all studies which don't. The assumptions presented in Section 2.3.1 would need to hold for such an approach. Notably, particular considerations include that:

- At least one of the studies meets the assumption of mean conditional exchangeability (thus, those study/ies need not be randomized, but can have no unmeasured confounding. This assumption is only guaranteed in expectation for randomized studies but substantive knowledge can be used to posit no unmeasured confounding in observational studies.)
- All studies meet Assumption 3 (i.e., one version of the intervention was applied in all of them, which precludes differential implementation across studies, differential measurement error of the outcome or intervention, etc.)
- The combined randomized and combined observational groups have some overlap with one another in their respective covariate distributions.
- The studies together are representative of the target population covariate distribution or can have their covariate distribution transformed such that $\mathcal{R}^* = \mathcal{R}^*_{\text{overlap}} \cup \mathcal{R}^*_{\text{obs-only}} \cup \mathcal{R}^*_{\text{RCT-only}}$ holds.

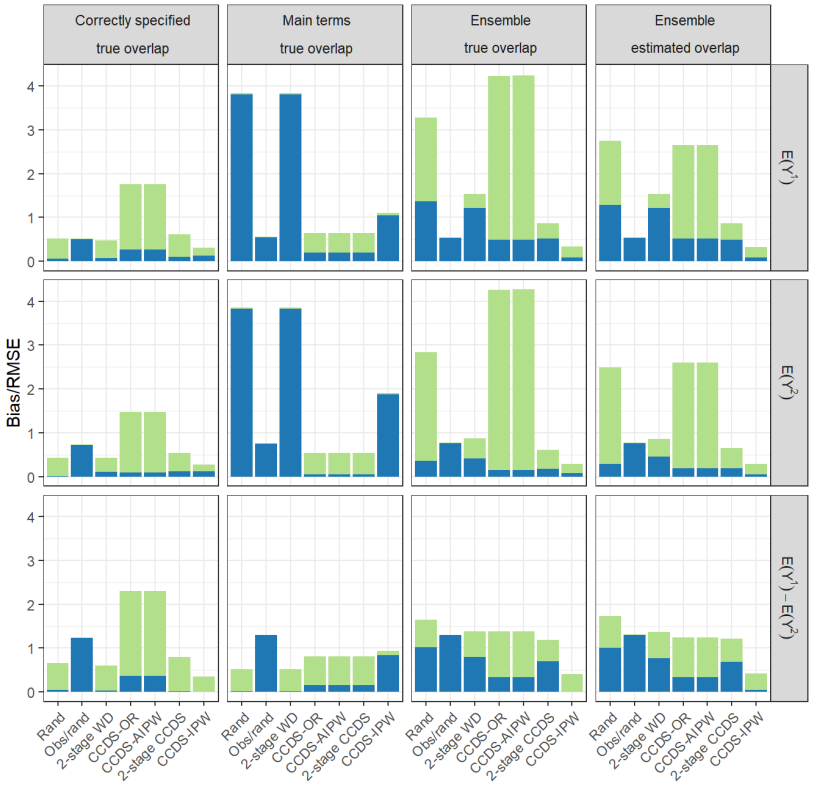
In such a collection of studies, after separating individuals into randomized and observational groups, analysis can proceed with the CCDS estimators presented in this paper.

**Alternative approaches for determining the region of overlap.** Our determination of the region of overlap is binary (observations are either in or out) and separate from the estimation step. The region of overlap could alternatively be specified to minimize mean squared error for estimating the PTSMs. Furthermore, rather than a sharp overlap region threshold, the degree of borrowing could be based on the similarity of randomized and observational data, as measured by the propensity score for selection or other covariate similarity metric.

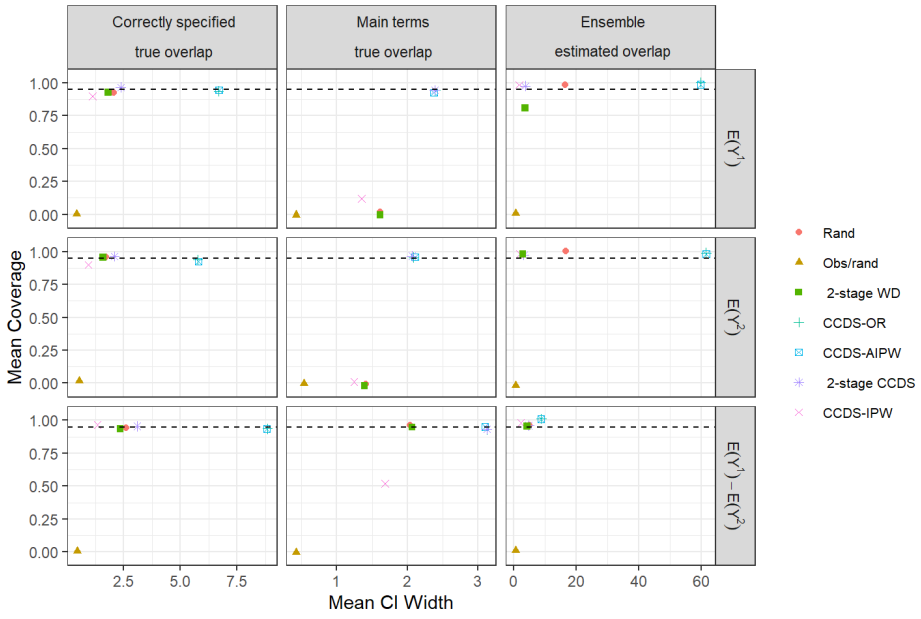
Furthermore, because we rely on an estimated propensity score for selection, determining whether points lie within the region of overlap is not only sensitive to hyperparameters $\alpha$

and $\beta$ but also relies on correct propensity model specification. One possible alternative is the convex hull approach presented by King and Zeng (2006), which avoids modeling the propensity score (and hence the possibility of model misspecification). Another alternative is described by Hill and Su (2013): using Bayesian additive regression trees (BART) to estimate "common causal support." It compares factual outcomes to BART-generated individual posterior distributions for each potential outcome to assess whether there exists sufficient information to make inference about that observation.

**Alternative approach for debiasing observational data.** Inspired by the representation of the potential outcome as $Y^{a,s}$, we can consider our goal as being the estimation of potential outcomes had everyone in the target population been in the randomized study ($S = 1$) in treatment group $a$. This suggests an alternative approach for bias estimation: (1) join randomized and observational data in the overlap region into a joint dataset, (2) fit a regression to the joint datasent, including $S$ as a covariate and accounting for interactions between $S$ and $\boldsymbol{X}$, then (3) predict counterfactual outcomes for the target sample or use observed outcomes for randomized data and predict counterfactual outcomes for just the observational data, setting $S = 1$ and $A = a$.



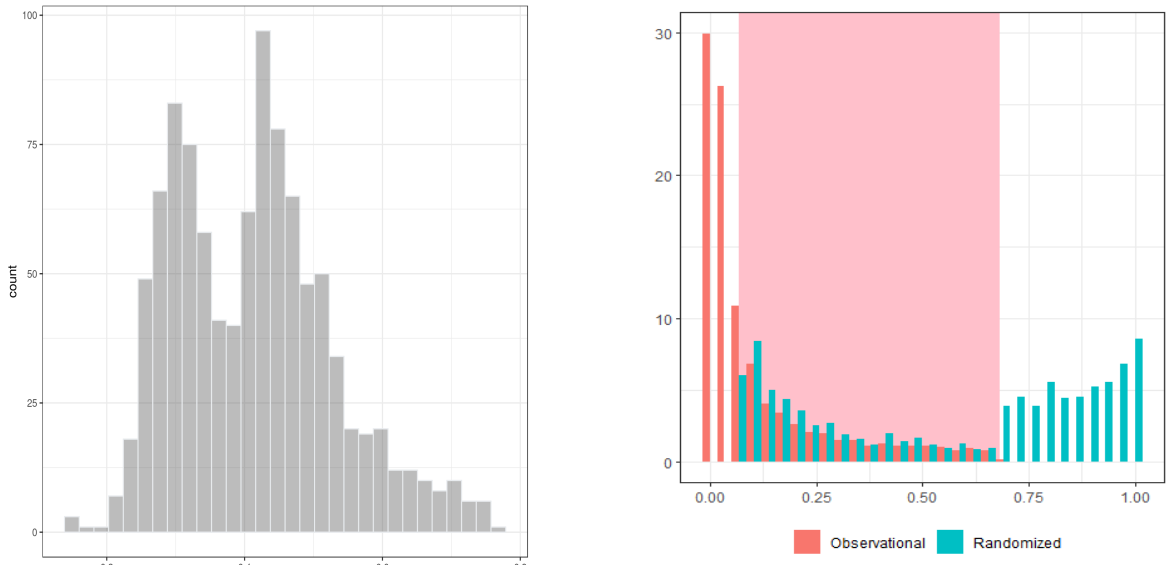

**(a)** *Bias and RMSE*




**(b)** *Coverage and CI Width*

**Figure (B.1)** *Performance across all estimators*

*Panel (a) depicts absolute bias is the darker portion of each bar; RMSE corresponds to the total bar size. In panel (b), the dashed line corresponds to the target coverage of 95%.*




**(a)** $\hat{P}(A = 1|S = 0, \mathbf{X}, \mathbf{U})$ **(b)** $\hat{P}(S = 1|X)$ *with true overlap region shaded*

**Figure (B.2)** *Estimated propensity scores for treatment and selection*

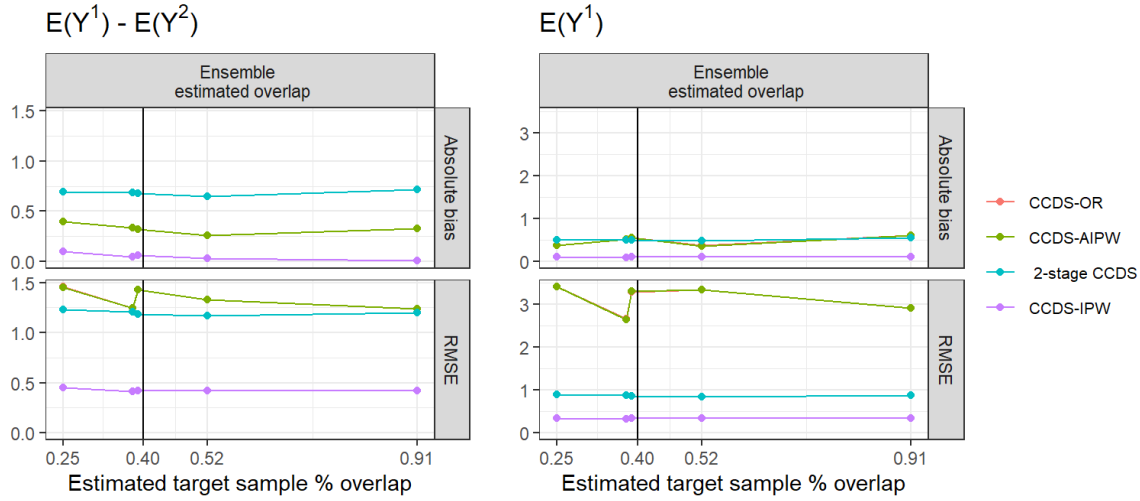



**Figure (B.3)** *Impact of different overlap region specifications on bias and RMSE*

*The true percent overlap is distinguished by a vertical black line.*

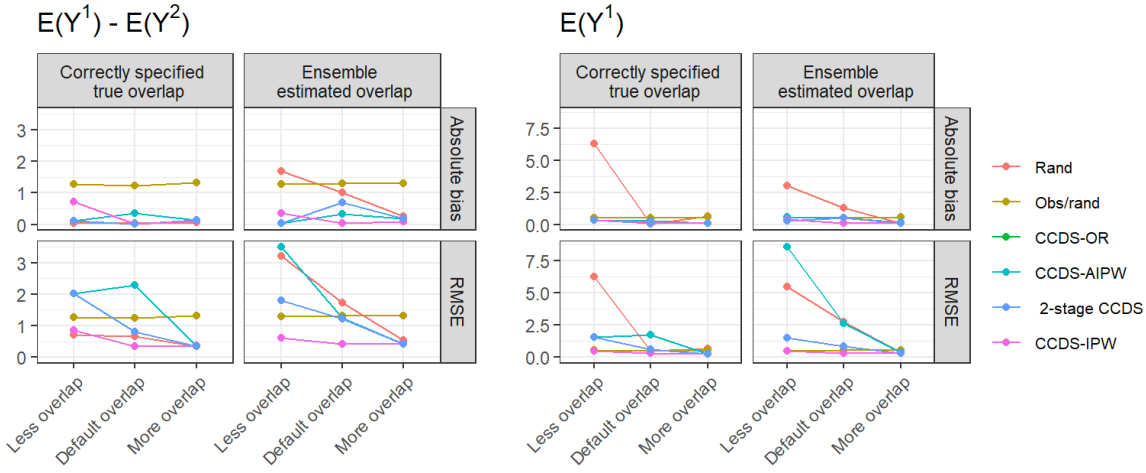



**Figure (B.4)** *Impact of degree of overlap (positivity of selection violation) on bias and RMSE*

*The settings examined include 13% (less overlap), 38% (default overlap), and 88% (more overlap) of observational data and 50% of randomized data in the overlap region.*

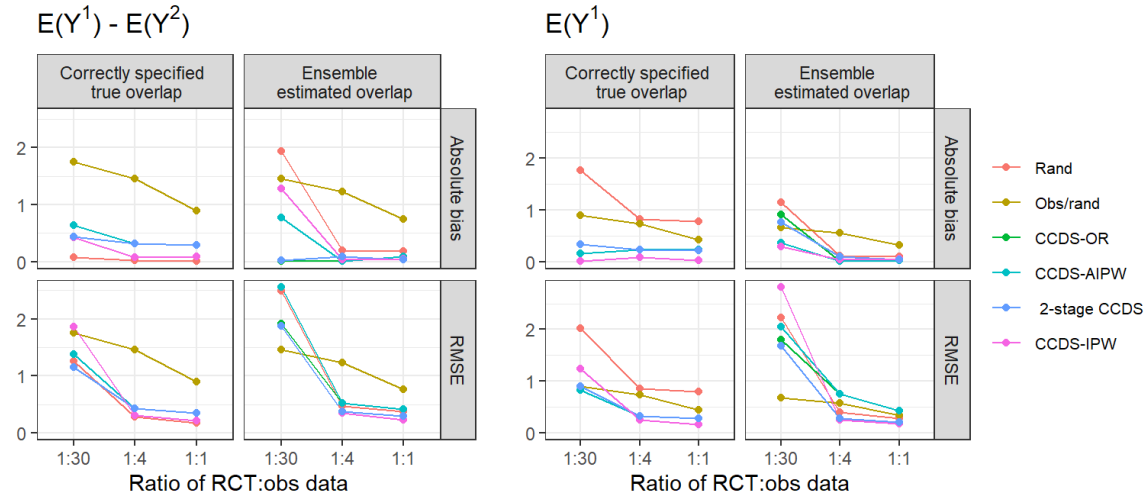



**Figure (B.5)** *Impact of different ratios of $n_{RCT} : n_{obs}$ on bias and RMSE*

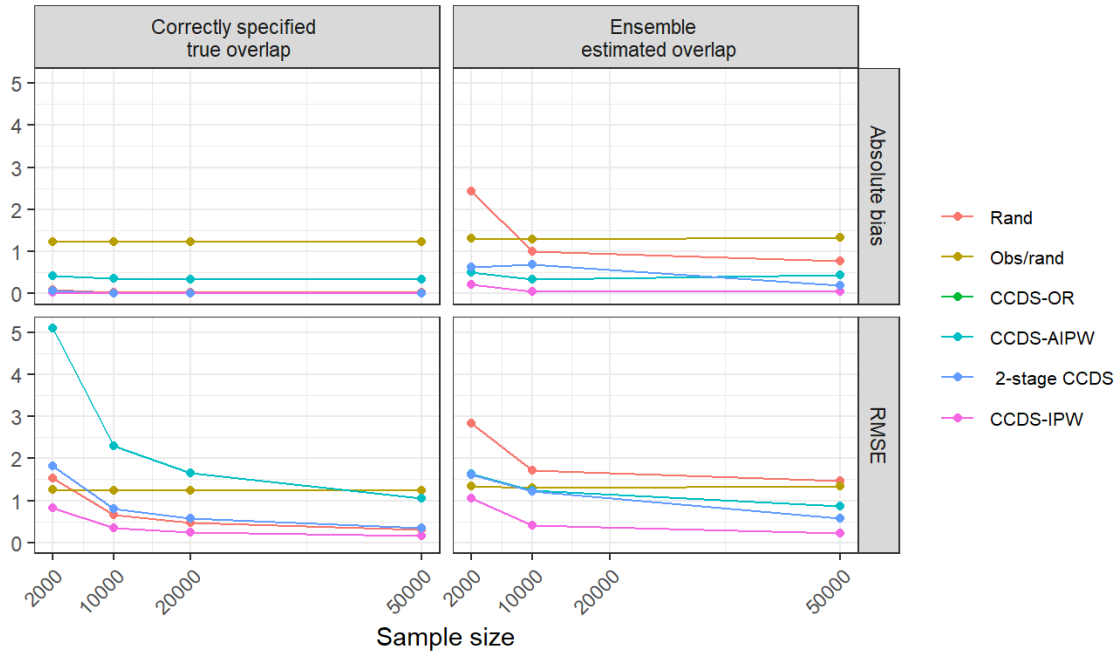



**Figure (B.6)** *Impact of n on bias and RMSE for PATE*

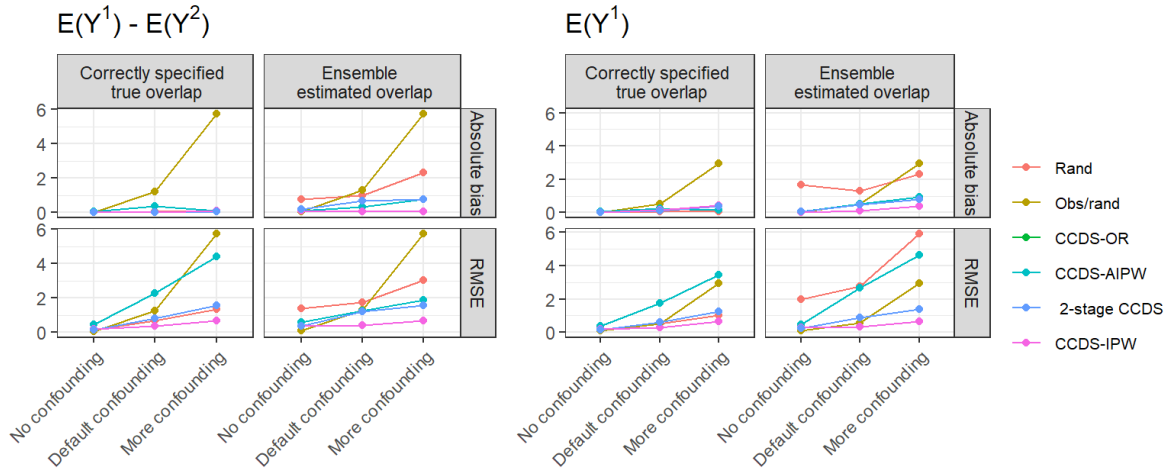



**Figure (B.7)** *Impact of unmeasured confounding on bias and RMSE*

*We examined settings with no unmeasured confounding, the default levels ($\beta_{AU} = 0.625$, $\beta_{YU} = 10$), and more confounding ($\beta_{AU} = 1.5$, $\beta_{YU} = 20$) for PATE.*

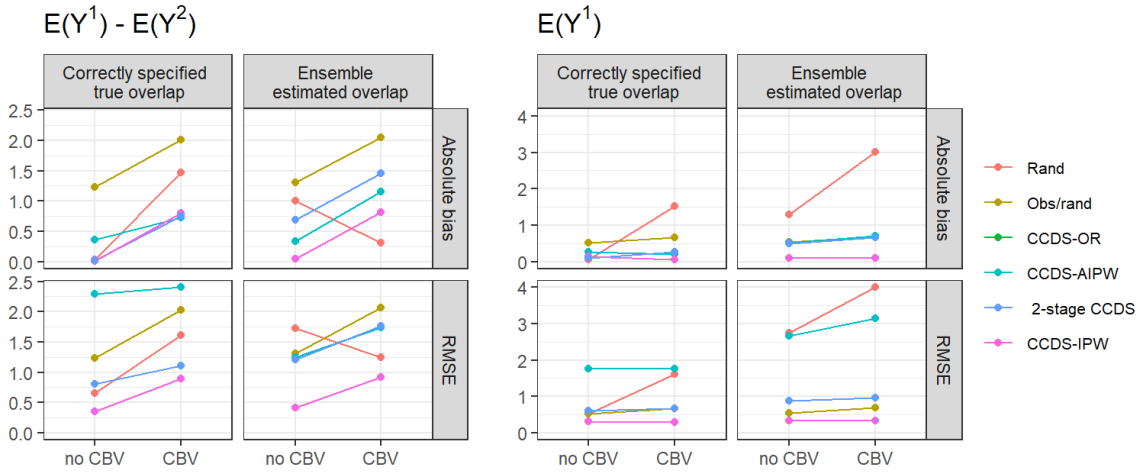
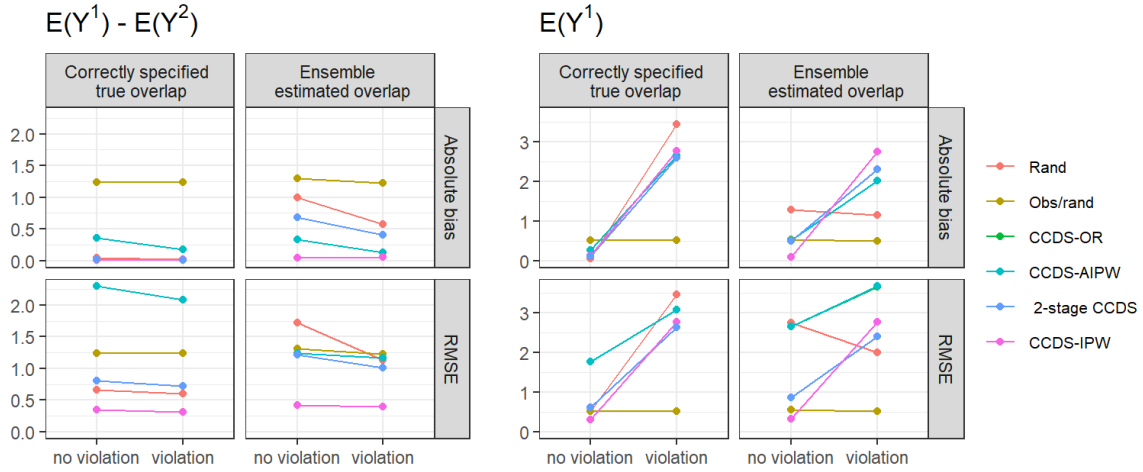
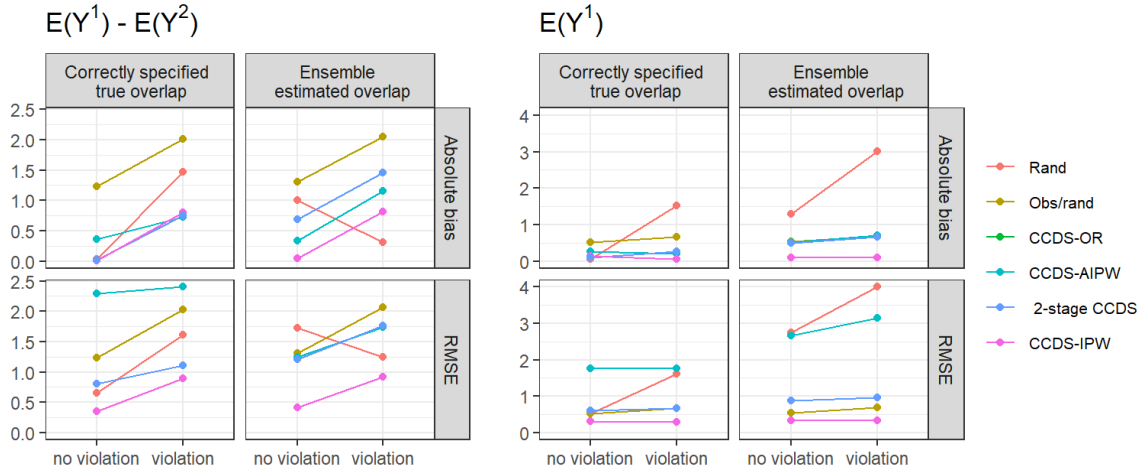



**Figure (B.8)** *Impact of constant conditional bias assumption violation on bias and RMSE*




**(a)** $S$ function of $U$




**(b)** $U$ function of $X_1$

**Figure (B.9)** *Impact of exchangeability of study selection assumption violation on bias and RMSE*

# Appendix C

# Appendix to Chapter 3

## C.1 Simulation data generating process

For the simulation, we generated $P(Y,A,X,S,R,V) = P(R)P(X|R)P(V|X)P(A,S|R,V,X)P(Y|A,X)$ as follows:

1. $R$ and $X_1,...,X_7$: The real data on which the simulation was based consisted of beneficiary- and practice-level claims data for the study and for all nationwide practices eligible for the scaled-up intervention. As most baseline covariates were categorical, we simulated baseline beneficiary- and practice-level covariates for the study region and non-study regions separately in the same proportions observed in the real data based on non-parametric assignment of practices to combinatoric cells describing possible baseline characteristic combinations (similar to that described in Lipman *et al.*).

2. $V|X \sim Binom(expit(\beta X))$: We fit a logistic propensity for volunteering regression to the real data study region practices (based on the pre-specified drivers of volunteering $X_4$ and $X_5$) to obtain $\beta$'s, which we used to generate a propensity for volunteering for all nationwide practices. Volunteering status was generated for each practice with probability corresponding to its propensity. Approximately 9% of practices volunteered, as was observed in the real data.

3. $A$ and $S$: Within the study region, practices which volunteered became the study treated group. To parallel the real data control selection process, study controls were then chosen from non-study regions via 4:1 matching. Controls are matched to study treated patients based on their propensity scores for treatment (estimated via logistic regression) using the `MatchIt` package in R (Ho *et al.*, 2011). Study region volunteering practices and matched controls from the non-study region formed the study sample.

4. $Y|A,X$: We generated the outcomes as $Y_t \sim 263 + c_0 + c_1 X_1 + c_2 X_1^2 + X_c - 45t + (6 + X_1 + 0.3X_1^2 + X_{\text{mod}})At$ with $t \in \{0,1\}$ indicating pre/post-intervention time; the outcome of interest being $Y = Y_{t=1} - Y_{t=0}$; $X_1$ corresponding to practice-level averages of beneficiary sickness levels, determined through a stochastic process described in Lipman *et al.* ($X_1$ is an unmeasured variable proxied by the measured variable age); $X_{\text{c}} = -238X_2 \times X_3 + 48log(X_4 + 0.1) + 96X_5 \times X_6 - 64X_7$; $X_{\text{mod}} = -40I(X_2 = 0) \times I(X_3 = 0) + 5X_1 \times I(X_2 = 1)$ where the amount of effect modification corresponded to that observed in the real data (an interaction between two covariates was similarly observed to modify treatment effects in the real data) and covariates comprising $X_{\text{mod}}$ were chosen for having large discrepancy between study and non-study regions (e.g. $P(X_1|R = 1) = 28\%$ vs. $P(X_1|R = 0) = 15\%$); and $c_0$, $c_1$, $c_2$ were unmeasured practice-level cost-multipliers that correspond to different practices incurring different costs for treating the same sickness levels (Lipman *et al.*).