Conditional cross-design synthesis estimators for generalizability in Medicaid

Irina Degtiar | Tim Layton | Jacob Wallace | Sherri Rose

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. However, few generalizability approaches exist for estimating causal quantities in a target population that is not well-represented by a randomized study but is reflected when additionally incorporating observational data. To generalize to a target population represented by a union of these data, we propose a novel class of conditional cross-design synthesis estimators that combine randomized and observational data, while addressing their estimates' respective biases—lack of overlap and unmeasured confounding. These methods enable estimating the causal effect of managed care plans on health care spending among Medicaid beneficiaries in New York City, which requires obtaining estimates for the 7% of beneficiaries randomized to a plan and 93% who choose a plan, who do not resemble randomized beneficiaries. Our new 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. Applying these methods, we find substantial heterogeneity in spending effects across managed care plans. This has major implications for our understanding of Medicaid, where this heterogeneity has previously been hidden. Additionally, we demonstrate that unmeasured confounding rather than lack of overlap poses a larger concern in this setting.

KEYWORDS
causal inference, external validity, selection bias, transportability, unmeasured confounding

1 | INTRODUCTION

Medicaid, administered by the Centers for Medicare & Medicaid Services, provides insurance for low-income and disabled Americans, covering one-fifth of all individuals in the United States (Centers for Medicare & Medicaid Services, 2020). New York City (NYC) Medicaid Managed Care (MMC) provides health insurance to most New York Medicaid beneficiaries. Appropriately managing these MMCs requires understanding their impact on health care spending. However, while MMCs differ in their spending, they also differ in the beneficiaries they attract. In NYC, beneficiaries who do not actively choose a health plan are randomized to one based on plans’ geographic eligibility. However, the 7% of beneficiaries who are randomized are not representative of the broader NYC Medicaid beneficiaries. Appropriately managing these MMCs requires understanding their impact on health care spending. However, while MMCs differ in their spending, they also differ in the beneficiaries they attract. In NYC, beneficiaries who do not actively choose a health plan are randomized to one based on plans’ geographic eligibility. However, the 7% of beneficiaries who are randomized are not representative of the broader NYC Medicaid beneficiaries. Appropriately managing these MMCs requires understanding their impact on health care spending. However, while MMCs differ in their spending, they also differ in the beneficiaries they attract. In NYC, beneficiaries who do not actively choose a health plan are randomized to one based on plans’ geographic eligibility. However, the 7% of beneficiaries who are randomized are not representative of the broader NYC Medicaid beneficiaries. Appropriately managing these MMCs requires understanding their impact on health care spending. However, while MMCs differ in their spending, they also differ in the beneficiaries they attract. In NYC, beneficiaries who do not actively choose a health plan are randomized to one based on plans’ geographic eligibility. However, the 7% of beneficiaries who are randomized are not representative of the broader NYC Medicaid beneficiaries.
population. Of the remaining 93% of enrollees who actively chose their health plan (i.e., observational beneficiaries), some are not well-represented by any randomized beneficiaries. Conversely, estimates from observational beneficiaries may be subject to unmeasured confounding from variables not captured in claims data driving plan choices and spending outcomes. Understanding MMC plans’ impact on health care spending in the full NYC Medicaid population requires addressing two challenges. The first is the randomized data’s lack of representativeness (i.e., its estimates’ lack of external validity: they may not reflect causal effects in the full NYC Medicaid target population nor represent subsets of the target population). The observational data estimates’ potential for unmeasured confounding is the second (i.e., its estimates’ lack of internal validity: they may be biased for the causal effects of the observational population). Existing studies have focused on the second challenge by working solely with the randomized subset of beneficiaries, but have not addressed the resulting estimates’ lack of external validity (Geruso et al., 2020). Similar challenges arise in settings ranging from clinical trials that exclude certain patient subsets (Lu et al., 2019; Prentice et al., 2006) to policy evaluation studies that aim to inform deployment in a different population (Kern et al., 2016).

Fusing randomized and observational data sources, ‘data fusion’ has the potential to address each data source’s shortcomings (Bareinboim & Pearl, 2016) and enable generalizing or transporting inference from a randomized study to a target population. However, few existing approaches leverage a combination of randomized and observational data to address the internal and external validity biases of each data source’s estimates (Degtiar & Rose, 2023). Approaches that do combine individual-level randomized and observational data face limitations when the target population (here, the full NYC Medicaid population) does not fully overlap with the randomized data and the observational study faces unmeasured confounding. Existing techniques extrapolate from the randomized data beyond their support (Kern et al., 2016), assume estimates from the included observational data have no unmeasured confounding (Kern et al., 2016; Lu et al., 2019), or allow for unmeasured confounding but include assumptions or setups that do not translate to the Medicaid setting.

This third set of approaches includes Bayesian calibrated risk-adjusted modeling, currently only deployed in the context of Cox proportional hazards survival regression, which necessitates a third external data source that has strong effect modifier overlap with both the randomized and observational data, unavailable for our setting (Henderson et al., 2017; Varadhan et al., 2016). Rosenman et al. (2023) assumed that treatment effects are identical within strata of effect modifiers, requiring data to be stratifiable into a limited number of effect modifier categories. Our Medicaid analysis requires hundreds of effect modifier strata (representing categorizations of beneficiaries’ baseline expenditures, age, neighborhood, and so forth), thus, this approach would be impractical. Cross-design synthesis methods stratify the data into randomized individuals, observational individuals eligible for the trial, and ineligible observational individuals; they then extrapolate the effect of the inclusion criteria to adjust for bias in the ineligible subset of the observational sample (Begg, 1992; Kaizar, 2011). In our Medicaid study, we have no knowledge of each beneficiary’s eligibility for the randomized group, so these cross-design synthesis methods are not applicable. A two-step regression approach by Kallus et al. (2018) assumes that the randomized covariate distribution is subsumed in the observational data distribution and does not directly extend to estimating population treatment-specific means (PTSMs; i.e., health plan-specific spending, our estimands of interest) rather than average treatment effects.

To obtain health plan-specific causal estimates of health care spending in the full NYC Medicaid population, we present a novel class of methods, which we refer to as conditional cross-design synthesis (CCDS) estimators. These approaches address many of the 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 used to ‘debias’ observational data estimates, under an assumption of constant conditional bias between data sources (which can be seen as an extension of the assumption made for cross-design synthesis; Kaizar 2011). Overlapping support is determined through an extension of an approach by Nethery et al. (2019). These techniques are robust to both unmeasured confounding when using observational data and positivity violations for selection into the randomized data, due to incorporating debiased observational data estimates. The estimators include outcome regression, two-step outcome regression, inverse probability weighting (IPW), and double robust augmented IPW 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.

Section 2 defines notation and the estimand of interest. Section 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 4 presents the novel CCDS estimators. We evaluate all estimators through a simulation study in Section 5 that highlights settings, where each CCDS estimator can be anticipated to perform well. Section 6 applies
these methods to our NYC Medicaid study examining the impact of managed care plans on health care spending. Section 7 concludes with a discussion.

2 NOTATION AND ESTIMAND

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 (RCT) and observational (obs) 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 the statistical model, namely, a collection of possible probability distributions. Each of these draws consists of an outcome \( Y \in \mathbb{R} \), the intervention \( A \in \mathcal{A} \), the vector of covariates \( X \in \mathcal{R} \in \mathbb{R}^d \), where \( \mathcal{R} \) is the region of support in the target population’s covariate distribution, and an indicator for selection into the randomized group \( S \in \{0, 1\} \). Thus, the observational unit for the target sample is \( O = (Y, A, X, S) \).

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 \( O_{\text{RCT}} = (Y, A, X, S = 1) \sim (Y, A, X|S = 1) \equiv P_{\text{RCT}} \). 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 \( O_{\text{obs}} = (Y, A, X, S = 0) \sim (Y, A, X|S = 0) \equiv P_{\text{obs}} \).

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 PTSMs: \( E(Y^a) \) for \( \forall a \in \mathcal{A} \), as have been explored in prior analyses with multiple unordered treatments (Rose & 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: \( E(Y^{a^d}) - E(Y^d) \)) are of interest.

2.3 Defining and determining overlap and nonoverlap regions

Randomized and observational covariate distributions differ: \( P(X|S = 1) \neq P(X|S = 0) \). Furthermore, a portion of the observational data may not be well-represented in the randomized data (and vice versa): \( P(X = x|S = 0) = 0 \) and \( P(X = x'|S = 1) = 0 \) for some \( x, x' \in \mathcal{R} \). However, there exists a region of overlap: 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) (Figure 1). Regions of nonoverlap therefore correspond to regions of the covariate distribution, where either only observational individuals (\( R_{\text{obs-only}} \)) or only randomized individuals (\( R_{\text{RCT-only}} \)) would be observed, that is, regions where units in one study population are not eligible to be in the other study population (e.g., \( R_{\text{obs-only}} = x \in \mathcal{R} : P(X = x|S = 1) > 0 \cap P(X = x|S = 0) = 0 \)). The target sample covariate distribution (\( R \)) can therefore be decomposed as: \( R = R_{\text{overlap}} \cup R_{\text{obs-only}} \cup R_{\text{RCT-only}} \) (Figure 1). Thus, \( R_{\text{obs}} = R_{\text{overlap}} \cup R_{\text{obs-only}} \) and \( R_{\text{RCT}} = R_{\text{overlap}} \cup R_{\text{RCT-only}} \). \( R_{\text{RCT-only}} \) and \( R_{\text{obs-only}} \) may be null sets. Let \( R \) be an indicator for being in the respective region, for example, \( R_{\text{overlap}} = 1 \) (membership in \( R_{\text{overlap}} \)).

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 are collected through survey sampling. In this case, through reweighting, one can map the randomized and observational study regions of covariate support, \( R_{\text{RCT}} \) and \( R_{\text{obs}} \).
into a transformation, \( R_{\text{RCT}} \rightarrow R_{\text{RCT}}^* \) and \( R_{\text{obs}} \rightarrow R_{\text{obs}}^* \), in which the decomposition above of \( R_{\text{RCT}}^* = R_{\text{overlap}}^* \cup R_{\text{obs-only}}^* \cup R_{\text{RCT-only}}^* \) holds. Note that this includes the possibility of the target population being represented by just the observational data (the transportability setting).

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, \( R_{\text{overlap}} \), we extend a data-driven approach for determining areas of treatment overlap based on propensity scores for treatment assignment (Nethery et al., 2019). We adopt a similar approach for the propensity score for study selection \( \pi_S = P(S|X) \), but on the logit scale to give more granularity to very low and high propensity scores. Namely, the estimated region of overlap, \( \hat{R}_{\text{overlap}} = \{ i : \hat{R}_{\text{overlap}}(i) = 1 \} \), consists of points in the logit of the propensity score for selection that have at least \( \hat{\beta} \) observations from each study group within an interval of size \( \alpha \) around that point. Throughout, hats, \( \hat{\cdot} \), correspond to estimators of their respective quantities. For high granularity, we use \( \alpha = 0.01 \times \text{range(logit}(\pi_S)) \) and \( \hat{\beta} = 0.01 \times \min(n_{\text{RCT}}, n_{\text{obs}}) \). In simulations, we assessed sensitivity to tuning parameters \( \alpha \) and \( \hat{\beta} \).

## 3.2 Relaxation of 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: \( E(Y^a|S = 1, A = a, X) = E(Y^a|S = 1, X) \) \( \forall a \in A \), and constant conditional bias in the observational group: \( E(Y^a|S = 0, A = a, R_{\text{overlap}} = 1, X) - E(Y^a|S = 1, A = a, R_{\text{overlap}} = 1, X) \) \( \forall a \in A, x \in R_{\text{obs}} \). (5b) Overlap between study samples: there exists a non-null set \( R_{\text{overlap}} \) such that \( P(S = s|X = x) > 0 \) \( \forall s \in S, x \in R_{\text{overlap}} \).

Assumption 1b corresponds to the same conditional bias relationship holding in \( R_{\text{overlap}} \) as \( R_{\text{obs}} \): \( b(a, x) = b(a, x|R_{\text{overlap}} = 1) \), \( \forall a \in A, x \in R_{\text{obs}} \), where \( b(a, x) \equiv E(Y^a|S = 0, A = a, X = x) - E(Y^a|S = 0, X = x) \). In our Medicaid application, we assume that this assumption is reasonable—the way socioeconomic status, the unmeasured confounder of primary concern, is related to measured covariates should not depend on whether a beneficiary would always choose their own health plan. See Web Appendix 1 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 \( X \) covariate distribution in the observational data (mean constant conditional bias): \( E_X[E(Y^a|S = 0, A = a, X) - E(Y^a|S = 1, A = a, X)] = 0 \) and \( E_X[E(Y^a|S = 0, A = a, R_{\text{overlap}} = 1, X) - E(Y^a|S = 1, A = a, R_{\text{overlap}} = 1, X)] = 0 \).

Assumption 1b states that the relationship between bias and measured covariates is unrelated to being in the overlap region, that is, that the distribution of unmeasured confounders does not differ between \( R_{\text{overlap}} \) and \( R_{\text{obs}} \), conditional on \( X \) and \( A \). This assumption is strictly weaker than the no unmeasured confounding assumption, which is nested within Assumption 1b: with no unmeasured confounding, \( b(a, x) = 0 \). Constant conditional bias can depend on measured covariates and the covariate distribution support is dichotomized into overlap and nonoverlap regions rather than predefined eligibility determining overlap region membership. We hence assume that the covariates \( X \) 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 extrapolate to and correct for bias.
in estimates from 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 unmeasured confounders 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 (Web Appendix 2).

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

Of note, the \( \mathbf{X} \) needed for Assumptions 1b and 2 and the \( \mathbf{X} \) needed for Assumption 4 and 5b may differ. As a result, the overlap region should exist with respect to outcome predictors but be large enough for Assumption 1b to hold. It is hence reasonable to use an \( \mathbf{X} \) matrix that contains all outcome predictors and confounders to assess all assumptions. Thus, as described earlier, our \( \mathbf{X} \) is the union of covariate sets needed for all assumptions to hold.

### 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:

\[
\psi_{\text{CCDS}}(a) = E_{\mathbf{X}|S=1}[E(Y|S=1, A=a, \mathbf{X})|S=1]P(S=1) + E_{\mathbf{X}|S=0}[E(Y|S=0, A=a, \mathbf{X})|S=0]P(S=0) - E_{\mathbf{X}|S=0}[E(Y|S=0, A=a, R_{\text{overlap}}=1, \mathbf{X}) - E(Y|S=1, A=a, R_{\text{overlap}}=1, \mathbf{X})|S=0]P(S=0).
\]

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

### 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, two-stage outcome regression, IPW, and double robust augmented IPW approaches.

#### 4.1 Conditional cross-design synthesis outcome regression estimator

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

\[
\hat{\psi}_{\text{CCDS-OR}}(a) = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}(S_i = 1, A_i = a, \mathbf{X}_i)I(S_i = 1)
+ \hat{Q}(S_i = 0, A_i = a, \mathbf{X}_i)I(S_i = 0)
- \left\{ \hat{Q}(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1, \mathbf{X}_i)
- \hat{Q}(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1, \mathbf{X}_i) \right\}I(S_i = 0),
\]

where \( \hat{R}_{\text{overlap}} \) is estimated as described in Section 2.3, \( \hat{Q}(S = s, A = a, \mathbf{X}) \) is an estimator for \( E(Y|S = s, A = a, \mathbf{X}) \), and \( \hat{Q}(S = s, A = a, \hat{R}_{\text{overlap}} = 1, \mathbf{X}) \) is an estimator of \( E(Y|S = s, A = a, R_{\text{overlap}} = 1, \mathbf{X}) \). 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 for this and other estimators and a conditional treatment-specific mean version of the CCDS-OR estimator are presented in Web Appendix 5.

#### 4.2 Two-stage conditional cross-design synthesis outcome regression estimator

The two-stage CCDS estimator replaces the debiasing term, the two outcome regressions in the third term, with a two-stage regression that explicitly models the bias as a function of covariates. This second-stage regression can be regularized or constrained to avoid overfitting to overlap region trends:

\[
\hat{\psi}_{2\text{-stage CCDS}}(a) = \frac{1}{n} \sum_{i=1}^{n} \hat{Q}(S_i = 1, A_i = a, \mathbf{X}_i)I(S_i = 1)
+ \hat{Q}(S_i = 0, A_i = a, \mathbf{X}_i)I(S_i = 0)
- \hat{b}(S_i = 1, a, \mathbf{X}_i)I(S_i = 0),
\]

with \( \hat{b}(S_i = 1, a, \mathbf{X}_i) \) estimated as follows. Stage (1): fit outcome regressions to the overlap region of the observational data and the overlap region of the randomized data. Using randomized overlap data, estimate an intermediate bias term, \( \hat{b}(S_i = 1, a, \mathbf{X}_i) = \hat{Q}(S = 1, A = a, \hat{R}_{\text{overlap},i} = 1, \mathbf{X}_i) - \hat{Q}(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1, \mathbf{X}_i)\).
difference in predicted counterfactual outcomes using the two regressions. As there is no bias in expectation from the randomized overlap data’s estimates, any estimated bias stems from the regression \( \hat{Q}(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) \). Stage (2): regress \( \hat{b}'(S_i = 1, a, X_i) \) from Stage (1) on covariates \( X_i \) in a weighted regression:

\[
\hat{b}'(S_i = 1, a, X_i) = \frac{\hat{w}_{\text{bias}}(X_i)}{\sum_{i=1}^{n} \hat{w}_{\text{bias}}(X_i)} \hat{g}(X_i)
\]

with \( \hat{w}_{\text{bias}}(X_i) = \frac{1(S_i = 1, \hat{R}_{\text{overlap},i} = 1) \hat{P}(S_i = 0|X_i)}{\hat{P}(\hat{R}_{\text{overlap},i} = 1|S_i = 1, X_i) \hat{P}(S_i = 1|X_i)} \),

and \( g(X) \) some regression function, such as linear regression. Finally, use this weighted regression to estimate the debiasing term for the observational data estimates, \( \hat{b}(S_i = 1, a, X_i) \).

This second stage focuses on the relationship between the bias estimates in the overlap region and measured covariates. 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

\[
P(S = 0) = E_X[P(S = 0|X)] = E_X[1(S = 1, R_{\text{overlap}} = 1)P(S = 0|X)/P(R_{\text{overlap}} = 1|S = 1, X)P(S = 1|X)].
\]

Reweighting will frequently not face issues under positivity of selection violations. This is because \( S = 1 \) data are used to estimate the bias term and therefore should not have many values close to zero for \( \hat{P}(S_i = 1|X_i) \), which is in the denominator of the weight. Thus, while weighting is not required in such a two-stage approach, the weights add robustness compared to an unweighted approach without common drawbacks of weighting, such as variance inflation due to unstable weights.

Web Appendix 6 presents a two-stage approach that does not restrict itself to the overlap region (two-stage whole data). This method 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 estimates.

### 4.3 Conditional cross-design synthesis inverse probability weighting estimator

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

\[
\hat{\psi}_{\text{CCDS-IPW}}(a) = \frac{n_{\text{RCT}}}{n} \left( \sum_{i=1}^{n} \hat{w}_i(S_i = 1, A_i = a, X_i) \right)^{-1} \sum_{i=1}^{n} \hat{w}_i(S_i = 1, A_i = a, X_i) Y_i
\]

\[
+ \frac{n_{\text{obs}}}{n} \left( \sum_{i=1}^{n} \hat{w}_i(S_i = 0, A_i = a, X_i) \right)^{-1} \sum_{i=1}^{n} \hat{w}_i(S_i = 0, A_i = a, X_i) Y_i
\]

\[
- \frac{n_{\text{obs}}}{n} \left( \sum_{i=1}^{n} \hat{w}_i(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) \right)^{-1} \sum_{i=1}^{n} \hat{w}_i(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) Y_i
\]

\[
- \left( \sum_{i=1}^{n} \hat{w}_i(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) \right)^{-1} \sum_{i=1}^{n} \hat{w}_i(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) Y_i
\]

where

\[
\hat{w}_1(S_i = 1, A_i = a, X_i) = \frac{1(S_i = 1, A_i = a)}{\hat{P}(A_i = a|S_i = 1, X_i)}, \hat{w}_2(S_i = 0, A_i = a, X_i) = \frac{1(S_i = 0, A_i = a)}{\hat{P}(A_i = a|S_i = 0, X_i)}
\]

\[
\hat{w}_3(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) = \frac{1(S_i = 0, A_i = a, \hat{R}_{\text{overlap},i} = 1)}{\hat{P}(\hat{R}_{\text{overlap},i} = 1|S_i = 0, X_i) \hat{P}(A_i = a|S_i = 0, \hat{R}_{\text{overlap},i} = 1, X_i)}
\]

\[
\hat{w}_4(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1, X_i) = \frac{1(S_i = 1, A_i = a, \hat{R}_{\text{overlap},i} = 1)}{\hat{P}(S_i = 1|X_i) \hat{P}(\hat{R}_{\text{overlap},i} = 1|S_i = 1, X_i) \hat{P}(A_i = a|S_i = 1, \hat{R}_{\text{overlap},i} = 1, X_i)}
\]
and $\hat{P}(\cdot)$'s correspond to estimators for their respective probabilities.

Here, positivity of selection violations will usually not lead to unstable weights because $\hat{P}(\tilde{R}_{\text{overlap},i} = 1|S_i = 1, X_i)\hat{P}(S_i = 1|X_i)$ 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 can face lack of efficiency and potentially unstable estimates, particularly from estimating the second bias term contribution weighted by $\hat{w}_4$. Its 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.

### 4.4 Conditional cross-design synthesis augmented inverse probability weighting estimator

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

$$
\hat{\varphi}_{\text{CCDS-AIPW}}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{w}_1(S_i = 1, A_i = a, X_i)}{\sum_{i=1}^{n} \hat{w}_1(S_i = 1, A_i = a, X_i)} \{ Y_i - \hat{Q}(S_i = 0, A_i = a, X_i) \} + \{ Y_i - \hat{Q}(S_i = 1, A_i = a, X_i) \} + \{ Y_i - \hat{Q}(S_i = 0, A_i = a, X_i) \} + \{ Y_i - \hat{Q}(S_i = 1, A_i = a, X_i) \} + 1(S_i = 0) \hat{Q}(S_i = 0, A_i = a, X_i)
$$

with $\hat{Q}(\cdot)$ and $\hat{w}_1(\cdot)$ through $\hat{w}_4(\cdot)$ as defined in Sections 4.1 and 4.3, respectively. CCDS-AIPW is a double robust estimator that is asymptotically efficient when the propensity and outcome regressions are estimated consistently. See Web Appendix 8 for a derivation of the efficient influence function.

### 4.5 Inference

Confidence intervals and standard errors in our machine-learning-based analyses were calculated using a non-parametric bootstrap (Efron & Tibshirani, 1994). By re-estimating overlap regions for each bootstrap draw, uncertainty in the overlap region membership was incorporated into PTSM inference. While a nonparametric bootstrap may lack consistency guarantees for machine learning estimators, new work suggests possible approaches with coverage guarantees (Coyle & van der Laan, 2018). When using parametric regressions, a sandwich variance approach can be used to derive sampling variance, following M-estimation theory. For the CCDS-AIPW estimator, the influence function (derived in Web Appendix 8) could alternatively be used. M-estimation and influence function-based approaches would not incorporate uncertainty around overlap region estimation.

### 4.6 Comparison estimators

No existing methods address both the overlap and unmeasured confounding challenges in our setting. We therefore compare against two simple approaches. The first (RCT 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):

$$
\hat{\varphi}_{\text{RCT}}(a) = n^{-1} \sum_{i=1}^{n} \hat{Q}(S_i = 1, A_i = a, X_i).
$$

This extrapolation may yield bias when the relationship between covariates and potential outcomes differs in $R_{\text{obs}}$ compared to $R_{\text{obs}}$ in a way that cannot be extrapolated from the randomized data.

The second (obs/RCT estimator) is similar to Kern et al. (2016) and Prentice et al. (2006), though those estimators fit one outcome regression to both randomized and observational data and estimate effects for either just the observational data or just the randomized data. The obs/RCT 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:

$$
\hat{\varphi}_{\text{obs/RCT}}(a) = n^{-1} \sum_{i=1}^{n} \hat{Q}(S_i = 1, A_i = a, X_i)1(S_i = 1) + \hat{Q}(S_i = 0, A_i = a, X_i)1(S_i = 0).
$$

This approach assumes there is no unmeasured
confounding in the observational data estimates. We used outcome regressions for RCT and obs/RCT estimators rather than approaches that incorporate propensities for selection, as the latter will result in denominators close to zero due to lack of overlap. In Web Appendix 9, we also compare against Lu et al. (2019)’s AIPW versions of the RCT and obs/RCT estimators.

5 | SIMULATION STUDIES

We designed a broad set of simulations to evaluate the finite sample performance of our novel CCDS estimators compared to alternative approaches for estimating PTSMs and the PATE, examining two treatment groups, $A \in \{1, 2\}$. We assessed these estimators’ performance in the presence of (1) complex data-generating mechanisms such that the randomized data do not extrapolate well outside their support for PTSMs (but extrapolate well for PATEs), (2) unmeasured confounding in the observational data estimates, and (3) positivity of selection violations, for which Assumptions 1b and 5b hold but Assumptions 1 and 5 do not. 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, ensembles, and alternative regression fits. In total, we examined 84 different data generating scenario $\times$ regression choice combinations. We present results for “correctly specified” settings, where regression functional forms were correctly specified using measured information but missing the unmeasured confounder, a misspecified “main terms” setting where cubic terms were missing from regression specifications, and “ensemble” settings where we fit an ensemble of algorithms. See Web Appendix 10 for further implementation details, the correspondence of our simulation design with identifiability assumptions, and descriptions of alternative data-generating mechanisms.

5.1 | Main findings

Results across different regression specifications highlight the estimators’ relative strengths and disadvantages (Figure 2, Web Appendix Table S2). The RCT estimator performed well when it could extrapolate well outside $R_{\text{RCT}}$ (the correctly specified setting and for PATEs, which were generated to extrapolate well from randomized data; Web Appendix 11.2), but it suffered from large bias in the main terms setting and large bias and variance when fitting ensembles. For the obs/RCT estimator, unmeasured confounding bias was present even when correctly specified regressions were fit, though the large observational sample size led to relatively low root mean square error (RMSE). Results from Lu et al. (2019)’s AIPW-based RCT and obs/RCT estimators were similar to those from the outcome regression-based RCT and obs/RCT estimators, although the bias of Lu et al. (2019)’s obs/RCT estimators tended to be larger. This may be due to the misspecification of both outcome and propensity regressions and using less data to fit each regression because of sample splitting (Web Appendix Figure S1).

Novel estimators addressed extrapolation and unmeasured confounding bias. 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) but suffered from large variability when fitting complex regressions in the overlap region (as observed in the correctly specified and ensemble settings). These estimators’ performance was almost identical at a sample size of $n = 10,000$. In comparison, the two-stage CCDS estimator decreased bias and variance (in the main terms setting, its estimates were identical to those of CCDS-OR due to linearity and additivity).

CCDS-IPW had the smallest bias and RMSE throughout all settings except when grossly misspecifying the propensity for selection, in the main terms setting. However, the estimator’s superior performance was due to the outcome model having more variability compared to propensity models; for example, the propensity for selection was deterministically assigned by $X_1$. With a less deterministic relationship and smaller propensity scores, CCDS-IPW’s bias increased (Web Appendix 10.10).

5.2 | Estimating overlap

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 (Web Appendix 10).

5.3 | Coverage

The obs/RCT 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). Thus, while the bias and RMSE of the CCDS estimators
may or may not decrease compared to the obs/RCT esti-
mator due to remnant estimation error from misspecifying 
regressions, which is particularly evident with ensemble 
approaches, the obs/RCT estimator’s poor coverage indi-
cates this can be a false indication of precision. The RCT 
estimator attained 0% coverage in the main terms set-
tings and had wider CI in the ensemble setting than most 
novel estimators.

5.4 Alternative data-generating mechanisms

CCDS estimator bias and RMSE shrunk with more over-
lap and increasing proportions of randomized data. As 
unmeasured confounding bias increased, there was no cor-
responding 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 unmea-
sured confounding. Violating the constant conditional bias 
assumption increased bias for the RCT and all CCDS esti-
mators, with CCDS estimators generally performing better 
than the RCT 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 individu-
als. Overall, results for the CCDS estimators were similar 
across alternative data-generating mechanisms. Further 
details can be found in Web Appendix 10.

6 MEDICAID STUDY

6.1 Methods

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 CCDS and comparison estimators. Health 
care spending was examined over 6 months on the log 
scale, as log(spending + 1), adjusting for baseline spend-
ing decile, age, documented sex, aid group, social security 
income status, neighborhood, and neighborhood poverty.
level. Geruso et al. (2020) provided further descriptions of the data. 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. Baseline spending was missing for 1% of beneficiaries and 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 (Polley et al., 2019) (glm, 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 $1 - 0.05/10$ level, for 10 plans.

### 6.1.1 Populations

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 (US$ 2796 vs. US$ 3052), were more likely to have a documented sex of female (59% vs. 40%), came from different neighborhoods and aid groups, and were less likely to be eligible for social security income (2% vs. 9%) (Table 1). Effect heterogeneity within the randomized data estimates was driven by aid group status, supplemental security income eligibility, and neighborhood effects; within the observational data estimates, 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.

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 of possible confounding. 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 were most concentrated (Web Appendix Figure S11). Using the conservative overlap hyperparameters $\alpha = 0.01 \times \text{range}(\text{logit}(\pi_0))$ and $\hat{\beta} = 0.01 \times n_{\text{RCT}}$ 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 of the need for CCDS estimators.

### 6.2 Causal estimates

Despite higher unadjusted mean spending in the randomized group, STSM causal estimates across all health plans were higher for the observational group (Figure 3). This discrepancy reflects both differences in measured characteristics and potential unmeasured confounding in the observational data estimates. Neither estimate aligned with PTSM estimates for the NYC Medicaid target population, which were consistently lower than randomized and observational STSMs (Figure 3 includes results for two CCDS estimators well suited to this setting; see Web Appendix Figure S12 for all estimators). Results remained similar when accounting for country–month–year correlation (Web Appendix II.1).

CCDS estimates support the RCT estimator’s PTSM results—consistent with the substantial overlap between randomized and observational covariate distributions—providing evidence that extrapolation is much less of a concern in this setting than unmeasured confounding bias. Furthermore, these CCDS and RCT PTSMs were widely discrepant from obs/RCT PTSMs and observational AIPW STSMs (the latter two largely aligned, as observational beneficiaries comprised 93% of the target sample). The wide discrepancy suggests a large amount of unmeasured confounding bias in the observational data estimates. While the RCT PTSM estimator provided reasonable estimates in this setting, the CCDS estimators were able to incorporate all data and did not rely on extrapolating spending estimates beyond the randomized data support. Moreover, the double robust CCDS-AIPW estimates were higher than RCT estimates (12.5%–16.7% difference in log spending) and confidence intervals were non-overlapping for all but plans D and I. Unlike in the simulations, CCDS-AIPW confidence intervals (with ensemble regressions) were not large, but 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 (Web Appendix Figure S12). CCDS-AIPW estimates sizable heterogeneity for health plan effects on spending—a 55% difference between the highest- and lowest-spending plans.

### 7 DISCUSSION

When observational and randomized data are both available, there is potential to overcome each data type’s limitations through their combination. This paper proposed a class of novel estimators that can surmount positivity of selection assumption violations in the randomized data and unmeasured confounding in the observational data estimates by using common support between
### Table 1: Characteristics of randomized and observational Medicaid groups.

| 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  |
| 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  |

Notes: The observational group refers to the 10% random subset of all observational beneficiaries. 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.

the data sources to remove unmeasured confounding bias.

In the NYC Medicaid data, solely extrapolating from the randomized data (RCT estimator) or assuming no unmeasured confounding in the observational data estimates (obs/RCT estimator) resulted in disparate estimates. CCDS provided evidence as to which target population estimates may be most reliable: its estimates aligned more closely with RCT estimates, suggesting that unmeasured confounding in the observational data estimates posed a bigger challenge than extrapolation from the randomized data. Substantively, our results reinforce prior findings based solely on randomized NYC Medicaid beneficiaries: managed care plan differences lead to substantial
variability in spending, highlighting the importance of supply-side mechanisms for addressing rising Medicaid expenditures (Geruso et al., 2020).

In general, choosing between the proposed outcome regression, propensity score, and double robust CCDS estimators requires trading off their varying strengths. CCDS-AIPW with linear regressions is a suitable default approach that recovers unbiased estimates even when the true data-generating process is more complex (though can be approximated by linear regressions). However, when fitting more complex regressions, it 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. With more complex regressions, the two-stage CCDS or CCDS-IPW may also be suitable, depending on whether the outcome relationship or the propensity for selection and treatment relationships can be better approximated. The two-stage approach improves performance compared to the CCDS-OR estimator by stabilizing initial estimates to alleviate overfitting to overlap region trends.

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 challenges to applying CCDS estimators in other settings may include 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 randomized and observational groups. However, even 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 degree of information borrowing to depend on 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 or estimating common causal support. Estimating causal quantities from randomized and observational data commonly faces challenges beyond those of positivity of selection violation and unmeasured confounding discussed here. These 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 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 are 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) and when randomized trials are embedded in electronic health record data, the observational data can provide information on patients included in and excluded from the trial. Policy evaluation studies can be combined with observational data from outside the evaluation geography to estimate scale-up impacts (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, or when, in the absence of a comprehensive target sample, a combination of randomized and observational studies may more fully represent the target population (Prentice et al., 2006). 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 estimates.

ACKNOWLEDGMENTS
This work was supported by NIH grants DP2MD012722, T32LM012411, and T32ES07142.

DATA AVAILABILITY STATEMENT
The data that support the findings of this paper are available from the Centers for Medicare and Medicaid Services. Restrictions apply to the availability of these data, which are not publicly available due to privacy.

OPEN RESEARCH BADGES
This article has earned Open Data and Open Materials badges. Data and materials are available at http://re3data.org/

ORCID
Irina Degtiar https://orcid.org/0000-0002-6056-2262
Sherri Rose https://orcid.org/0000-0002-9076-8472

REFERENCES
Bareinboim, E. & Pearl, J. (2016) Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27), 7345–7352.
Begg, C.B. (1992) Cross design synthesis: a new strategy for medical effectiveness research. United States General Accounting Office, GAO/PEMD-92-18.
Centers for Medicare & Medicaid Services (2020) Medicaid facts and figures. https://www.cms.gov/newsroom/fact-sheets/medicaid-facts-and-figures.
Coyle, J. & van der Laan, M.J. (2018) Targeted bootstrap. In: M.J. van der Laan & S. Rose (Eds.) Targeted learning in data science: causal inference for complex longitudinal studies, Springer Series in Statistics. Cham: Springer International Publishing, pp. 523–539.
Degtiar, I. & Rose, S. (2023) A review of generalizability and transportability. Annual Review of Statistics and its Application, 10(1), 501–524.
Efron, B. & Tibshirani, R.J. (1994) An introduction to the bootstrap. Hall/Monographs on Statistics & Applied Probability. Boca Raton, FL: CRC Press.
Geruso, M., Layton, T.J. & Wallace, J. (2020) Are all managed care plans created equal? Evidence from random plan assignment in Medicaid. Technical Report w27762, National Bureau of Economic Research.
Henderson, N.C., Varadhan, R. & 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.
Kaizar, E.E. (2011) Estimating treatment effect via simple cross design synthesis. Statistics in Medicine, 30(25), 2986–3009.
Kallus, N., Puli, A.M. & Shalit, U. (2018) Removing hidden confounding by experimental grounding. Advances in neural information processing systems, vol. 31. Curran Associates, Inc., Red Hook, NY, United States.
Kern, H.L., Stuart, E.A., Hill, J. & Green, D.P. (2016) Assessing methods for generalizing experimental impact estimates to target
populations. Journal of Research on Educational Effectiveness, 9(1), 103–127.

Lu, Y., Scharfstein, D.O., Brooks, M.M., Quach, K. & Kennedy, E.H. (2019) Causal inference for comprehensive cohort studies. https://arxiv.org/abs/1910.03531

Nethery, R.C., Mealli, F. & Dominici, F. (2019) Estimating population average causal effects in the presence of non-overlap: the effect of natural gas compressor station exposure on cancer mortality. The Annals of Applied Statistics, 13(2), 1242–1267.

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

Prentice, R.L., Langer, R.D., Stefanick, M.L., Howard, B.V., Pettinger, M., Anderson, G.L., et al. (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.

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

Rose, S. & 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.T.R., Basse, G., Owen, A.B. & Baiocchi, M. (2023) Combining observational and experimental datasets using shrinkage estimators. Biometrics.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. & 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.

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

Varadhan, R., Henderson, N.C. & 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.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 3–6 are available with this paper at the Biometrics website on Wiley Online Library. R code used for the simulation study in Section 4 is available at the Biometrics website and at https://github.com/idegtiar1/CCDS.

Figure S1: Performance across all estimators.

Table S1: Population and sample true potential outcome means and means observed in each treatment group in simulation study.

Figure S2: Estimated propensity scores for treatment and selection in simulation study.

Table S2: Simulation results across 2,000 simulation iterations and 1,000 bootstrap replications.

Table S3: Overlap region specifications.

Figure S3: Impact of different overlap region specifications on bias and RMSE.

Figure S4: Impact of degree of overlap (positivity of selection violation) on bias and RMSE.

Figure S5: Impact of different ratios of n_{RCT} : n_{obs} on bias and RMSE.

Figure S6: Impact of n on bias and RMSE for PATE.

Figure S7: Impact of unmeasured confounding on bias and RMSE.

Figure S8: Impact of constant conditional bias assumption violation on bias and RMSE.

Figure S9: Impact of exchangeability of study selection assumption violation on bias and RMSE.

Figure S10: Impact of probabilistic π_S and smaller P(A = 1) on bias and RMSE.

Table S4: STSMs and PTSMs across Medicaid data analyses with differing correlation procedures for country–month–year (CMY).

Figure S11: Propensity for selection into the randomized group in the Medicaid data.

Figure S12: STSMs and PTSMs across health plans for all estimators, with 95% confidence intervals multiplicity-adjusted with the Bonferroni correction.

Data S1

How to cite this article: Degtiar, I., Layton, T., Wallace, J. & Rose, S. (2023) Conditional cross-design synthesis estimators for generalizability in Medicaid. Biometrics, 79, 3859–3872. https://doi.org/10.1111/biom.13863