# Generalizing Clinical Trials with Convex Hulls

**Eric V. Strobl** ERIC.STROBL@VUMC.ORG and **Thomas A. Lasko** TOM.LASKO@VUMC.ORG
*Vanderbilt University Medical Center*

**Editors:** Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

## Abstract

Randomized clinical trials eliminate confounding but impose *strict* exclusion criteria that limit recruitment to a subset of the population. Observational datasets are more inclusive but suffer from confounding – often providing overly optimistic estimates of treatment response over time due to partially optimized physician prescribing patterns. We therefore assume that the unconfounded treatment response lies somewhere in-between the observational estimate *before* and the observational estimate *after* treatment assignment. This assumption allows us to extrapolate results from exclusive trials to the broader population by analyzing observational and trial data simultaneously using an algorithm called Optimum in Convex Hulls (OCH). OCH represents the treatment effect either in terms of convex hulls of conditional *expectations* or convex hulls (also known as mixtures) of conditional *densities*. The algorithm first learns the component expectations or densities using the observational data and then learns the linear mixing coefficients using trial data in order to approximate the true treatment effect; theory importantly explains *why* this linear combination should hold. OCH estimates the treatment effect in terms both expectations and densities with state of the art accuracy.

**Keywords:** Causal inference, cross-design synthesis, randomized clinical trial, observational data

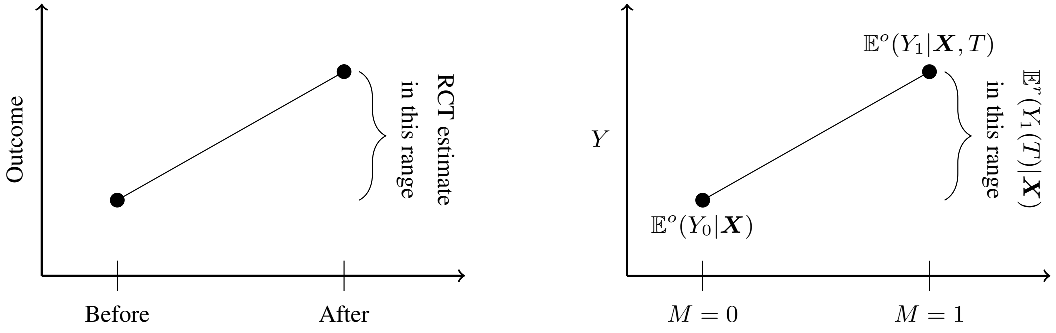
## 1. Introduction

Randomized clinical trials (RCTs) are the gold standard for inferring causal effects of treatment. RCTs eliminate confounding by randomizing treatment assignment. Randomization however imposes ethical and practical limitations that necessitate *strict* inclusion and exclusion criteria in practice. As a result, trials impose selection bias by limiting entry to a select sub-population. Inferences made with RCTs can thus fail to generalize to everyone seeking help.

Observational datasets, on the other hand, do not randomize treatment assignment. As a result, they suffer from confounding but impose much milder criteria for entry into the study. Inferences made with observational data generalize to the broader population but may not recover the true causal effect no matter how complicated the fit.

Stated succinctly, observational datasets are inclusive but confounded, whereas RCTs are exclusive but unconfounded. We thus propose to analyze RCT and observational data *simultaneously* in order to eliminate both confounding and selection bias. To this end, we exploit the following observation:

> Physicians are like reinforcement learning agents; they identify sub-populations of patients that appear to respond well to a given treatment *over time*, and then prescribe that treatment more often to those patients. The improvement in patient outcomes from before to after treatment assignment is therefore confounded in observational data and usually too large.


Outcome
RCT estimate
in this range
Before
After
$Y$
$\mathbb{E}^o(Y_1|\boldsymbol{X},T)$
$\mathbb{E}^r(Y_1(T)|\boldsymbol{X})$
in this range
$\mathbb{E}^o(Y_0|\boldsymbol{X})$
$M=0$
$M=1$

(a)

(b)

Figure 1: Illustration of the main assumption used in this paper. Lines denote the change in response to a given treatment over time in the observational data. (a) The RCT estimate lies in-between the observational estimates seen before and after treatment administration. (b) The same figure as (a) but with potential outcomes notation.

Such partially optimized physician prescribing patterns exist before an RCT is ever conducted. Beach et al. (2017) for instance even recommend entire sequences of treatments for different sub-populations based on evidence derived almost exclusively from case reports.

Physicians however also want to isolate the treatment effect from confounding bias using a rigorous RCT. RCTs ignore sub-populations by randomizing treatment assignment, so we expect the RCT estimate of patient outcomes to lie somewhere in-between the observational study estimates seen before and after treatment administration (Figure 1 (a)). For example, observational studies of lithium suggested that the medication decreases suicide attempts by a factor of two to three (Goodwin et al., 2003; Hayes et al., 2016). Physicians often prescribe lithium to chronically suicidal patients in the hopes that the medicine will decrease suicide attempts *in the future*. Large double blinded RCTs of lithium in broader populations replicated the decrease in suicide attempts over time but at a much smaller magnitude (Lauterbach et al., 2008; Oquendo et al., 2011; Lit, 2021). Observational studies therefore yielded an effect size that was in the same direction but too large.

We will convert the above observation into a precise assumption after reviewing the potential outcomes framework. We also develop an algorithm called Optimum in Convex Hulls (OCH) that exploits the highlighted observation by analyzing observational and trial data simultaneously. OCH estimates the treatment effect across the entire population in terms of the difference of two convex hulls. The algorithm can recover the treatment effect in terms of conditional expectations or even conditional densities with state of the art accuracy.

## 2. Potential Outcomes

We assume binary treatment assignment denoted by the random variable $T$ with $T=0$ or $T=1$, and $\mathbb{P}(T)>0$. We also adopt the potential outcomes framework, where we assume the existence of two potential outcomes $Y(0)$ and $Y(1)$ for all patients. However, we can only observe the single outcome $Y(t)$ for the subject assigned to $T=t$.

Let $\boldsymbol{X}$ denote the set of patient covariates measured prior to treatment assignment. Randomizing treatment assignment over the *entire* population ensures that we have $\{Y(0),Y(1)\} \perp\!\!\!\perp T|\boldsymbol{X}$ and

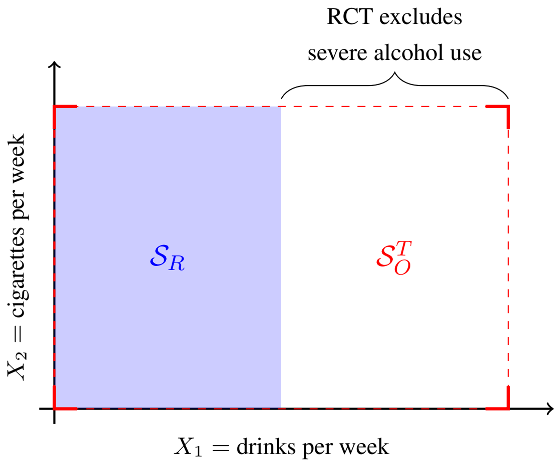


Figure 2: Example of an RCT strictly excluding patients who drink alcohol excessively. The support $\mathcal{S}_O^T$ is outlined with a dashed red line and $\mathcal{S}_R$ is shaded in blue; notice that we have $\mathcal{S}_R \subset \mathcal{S}_O^T$.

$\boldsymbol{X} \perp\!\!\!\perp T$, so that treatments are given regardless of potential outcomes and patient characteristics. However, patients and clinicians may find randomization disturbing because they cannot ensure optimal treatment assignment. RCTs therefore impose *strict* exclusion criteria[1] in practice based on pre-treatment covariates $\boldsymbol{X}$ that limit entry into the study. For example, an RCT examining the effects of anti-depressants may exclude patients with severe alcohol use, since these patients are more likely to benefit from inpatient detoxification. Without loss of generality, let $S$ denote a binary random variable representing selection bias (not a study indicator); the variable takes on a value of one if a patient is included in the RCT and zero otherwise.

RCTs impose selection bias, but they still randomize treatment assignment *among the recruited* so that we have:

**Assumption 1.** *$\{Y(0), Y(1)\} \perp\!\!\!\perp T | \{\boldsymbol{X}, S = 1\}$ and $\boldsymbol{X} \perp\!\!\!\perp T | S = 1$ in the RCT distribution.*

Trials therefore eliminate confounding but only for the recruited sub-group. These independence relations do not hold in the observational distribution in general. RCTs then sample patient covariates from the distribution $\mathbb{P}(\boldsymbol{X}|T, S = 1) = \mathbb{P}(\boldsymbol{X}|S = 1)$ with support $\mathcal{S}_R$, while observational studies sample from the distribution $\mathbb{P}(\boldsymbol{X}|T) = \mathbb{E}_{S|T}[\mathbb{P}(\boldsymbol{X}|T, S)]$ with support $\mathcal{S}_O^T$. We also have:

**Assumption 2.** *$\mathcal{S}_R \subseteq \mathcal{S}_O^T$,*

since RCTs impose selection bias with exclusion criteria whereas observational datasets do not. We provide an illustration in Figure 2 using the anti-depressant example, where patients with severe alcohol use are excluded from the RCT. We would like to draw conclusions about this sub-population as well because some of these patients drink large amounts of alcohol to cope with depression. Any procedure that makes inferences on $\mathcal{S}_R$ must therefore *extrapolate* to $\mathcal{S}_O^T \setminus \mathcal{S}_R$ in order to generalize to the broader population.

Since RCTs construct their exclusion criteria based on $\boldsymbol{X}$, we also frequently have:

[1] RCTs also utilize inclusion criteria, which we can convert to exclusionary ones by logical negation.

**Assumption 3.** $\{Y(0), Y(1)\} \perp\!\!\!\perp S|\boldsymbol{X}$ *in the RCT distribution.*

For instance, $\boldsymbol{X}$ may include amount of alcohol use in the aforementioned example, and patients who exceed a certain threshold are excluded from the study. $S$ therefore provides no additional information on a potential outcome given $\boldsymbol{X}$: $\mathbb{P}(Y(T)|\boldsymbol{X}, S) = \mathbb{P}(Y(T)|\boldsymbol{X})$. Note that this assumption only applies to $\mathcal{S}_R$. We provide an overview of Assumptions 1-3 in terms of graphical models in the Appendix 7.1.

In summary, RCTs impose selection bias (Assumption 2), often using predetermined criteria (Assumption 3), but eliminate confounding among the recruited (Assumption 1). In contrast, observational studies eliminate selection bias (Assumption 2) but introduce confounding. We therefore focus on eliminating both selection bias and confounding by analyzing RCT and observational data simultaneously. We in particular aim to estimate the *conditional average treatment effect* (CATE) given by $g(\boldsymbol{X}) \triangleq \mathbb{E}(Y(1)|\boldsymbol{X}) - \mathbb{E}(Y(0)|\boldsymbol{X})$ for everyone, or on $\mathcal{S}_O = \mathcal{S}_O^1 \cap \mathcal{S}_O^0$. The CATE corresponds to a difference of two conditional expectations, where we condition on $\boldsymbol{X}$ in order to identify patient-specific treatment effects in the spirit of precision medicine.

The CATE unfortunately only provides a point estimate for each patient and therefore does not take into account the uncertainty in the outcome value. Patients understand this uncertainty and frequently want to know the probabilities associated with all possible outcomes. We therefore also seek to recover *conditional densities of treatment effect* (CDTE), or $p(Y(1)|\boldsymbol{X})$ and $p(Y(0)|\boldsymbol{X})$ on $\mathcal{S}_O$. The CDTE summarizes the probabilities associated with all possible outcome values, i.e. $p(Y(1) = y|\boldsymbol{X})$ and $p(Y(0) = y|\boldsymbol{X})$ for all possible $y$ (Figure 6 in Appedix 7.2). We will recover the CATE and CDTE even for patients excluded from the trial by analyzing RCT and observational data simultaneously.

## 3. Main Assumptions

We now introduce the main assumptions used in this paper. Unlike Assumptions 1-3, the assumptions proposed in this section impose functional restrictions that may not map onto changes in graphical structure. We will consider two time steps: before and after treatment assignment, corresponding to the binary random variable $M = 0$ and $M = 1$, respectively. We use the notation $Y_M(T)$ to denote the potential outcome with treatment assignment $T$ at time step $M$. Treatment cannot causally affect the outcome before treatment assignment, or at $M = 0$.

### 3.1. Conditional Expectations

Let $Y_M = TY_M(1) + (1 - T)Y_M(0)$ denote the observed potential outcome. We can then write the conditional expectation in the observational distribution as $\mathbb{E}^o(Y_M(T)|\boldsymbol{X}, T) = \mathbb{E}^o(Y_M|\boldsymbol{X}, T)$ and that in the ideal RCT distribution, where everyone is randomized, as $\mathbb{E}^r(Y_M(T)|\boldsymbol{X})$ because $Y_M(T) \perp\!\!\!\perp T|\boldsymbol{X}$ in this case. The superscripts emphasize the observational and RCT distributions.

We in general have $\mathbb{E}^o(Y_0|\boldsymbol{X}, T = 0) \neq \mathbb{E}^o(Y_0|\boldsymbol{X}, T = 1)$ in observational studies even before treatment assignment due to confounding. Physicians may for instance choose to give a stronger medication $T = 1$ to sicker patients. Sicker patients have worse outcomes even before treatment assignment as reflected by the random variable $Y_0(1)$. However, the quantity:

$$\mathbb{E}^o(Y_0|\boldsymbol{X}) = \sum_{t \in \{0,1\}} \mathbb{E}^o(Y_0|\boldsymbol{X}, T = t)\mathbb{P}^o(T = t|\boldsymbol{X}) \tag{1}$$

provides a baseline value for *all* patients with covariates $\boldsymbol{X}$ before treatment has time to take effect, or regardless of whether $T = 1$ or $T = 0$ at time step $M = 0$.

As alluded to in the introduction, observational studies tend to produce overly large estimates of treatment response over time, so the treatment response in the RCT should lie somewhere in-between the observational estimate before and after treatment assignment. We now more explicitly assume:

**Assumption 4.** $\mathbb{E}^r(Y_1(T)|\boldsymbol{X}) = \mathbb{E}^o(Y_1|\boldsymbol{X}, T)\mu_T + \mathbb{E}^o(Y_0|\boldsymbol{X})(1 - \mu_T)$ *where* $\mu_T \in [0, 1]$.

The above equality holds for $T = 0$ and $T = 1$, and we may have $\mu_0 \neq \mu_1$. In other words, $\mathbb{E}^r(Y_1(T)|\boldsymbol{X})$ in the RCT distribution lies in the convex hull of $\mathbb{E}^o(Y_0|\boldsymbol{X})$ and $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)$ for before and after treatment assignment, respectively (Figure 1 (b)).

Physicians can detect effective treatments in sub-populations by observing the change in their patients before and after treatment administration. We therefore do not expect treatment in the RCT to have an opposite effect over time, or obtain $\mu_T < 0$. We similarly do not expect randomization to outperform the intelligent selection of treatment by trained professionals, or obtain $\mu_T > 1$. However, clinicians then start prescribing certain medications to select patients based on baseline characteristics $\boldsymbol{X}$, so confounding may account for most of the improvement in the mean outcome seen in observational data – implying that $\mu_T \leq 1$. Assumption 4 formalizes these ideas within the potential outcomes framework. We will also see that enforcing $\mu_T \in [0, 1]$, as opposed to allowing $\mu_T \in \mathbb{R}$, increases the robustness of the proposed algorithm to stricter exclusion criteria in Section 5.2.

We do *not* simply assume $g(\boldsymbol{X}) = \gamma[\mathbb{E}^o(Y_1|\boldsymbol{X}, T = 1) - \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 0)]$ with $\gamma \in [0, 1]$, so that the CATE either matches or underperforms the observational estimate at $M = 1$. Assumption 4 uses the magnitude of change within each treatment in congruence with how physicians choose a treatment that optimizes patient outcome over time like reinforcement learning agents. The CATE can have a larger magnitude than the observational estimate at $M = 1$ under Assumption 4; see for example Figure 7 in Appendix 7.3.

Assumption 4 similarly differs from the Manski bounds (Manski, 1989), where $\mathbb{E}^r(Y_1(T)|\boldsymbol{X})$ is bounded above by $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)\mathbb{P}^o(T|\boldsymbol{X}) + \mathbb{P}^o(\neg T|\boldsymbol{X})$ and below by $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)\mathbb{P}^o(T|\boldsymbol{X})$ when $Y_1(T) \in [0, 1]$. Assumption 4 again bounds changes *over time*, whereas the Manski bounds only apply to $M = 1$.

Finally, Assumption 4 differs from the parallel slopes assumption used in conditional Difference in Differences (Card and Krueger, 1993; Abadie, 2005). The parallel slopes assumption enforces the equality $\mathbb{E}^o(Y_1(0)|\boldsymbol{X}, T = 1) - \mathbb{E}^o(Y_0(0)|\boldsymbol{X}, T = 1) = \mathbb{E}^o(Y_1(0)|\boldsymbol{X}, T = 0) - \mathbb{E}^o(Y_0(0)|\boldsymbol{X}, T = 0)$, or *equalizes* the slopes of change across treatment assignments. Difference in Differences then identifies the conditional average treatment effect *on the treated*. In contrast, Assumption 4 bounds the RCT estimate using the slopes of change in the observational distribution. We will exploit Assumption 4 in Section 4 to identify the CATE irrespective of treatment assignment.

### 3.2. Conditional Densities

We can extend Assumption 4 to conditional densities as follows:

**Assumption 5.** $p^r(Y_1(T)|\boldsymbol{X}) = \underbrace{p^o(Y_1|\boldsymbol{X}, T)}_{(1)}\mu_T + \underbrace{p^o(Y_0|\boldsymbol{X})}_{(2)}(1 - \mu_T)$ *where* $\mu_T \in [0, 1]$, so that we no longer restrict ourselves to a convex hull of the expectations but to a mixture of two conditional densities. Clearly Assumption 5 implies Assumption 4 but not vice versa. Assumption 5 also implies that we can decompose the density $p^r(Y_1(T)|\boldsymbol{X})$ into two groups of patients: (1) those who respond to treatment just like in the observational dataset, and (2) those who do not respond. The coefficients $\mu_T$ and $1-\mu_T$ represent learnable parameters that correspond to the unknown proportion of patients who satisfy (1) and (2), respectively.

## 4. Optimum in Convex Hulls

We want to isolate the effect of treatment from the confounding bias introduced by partially optimized physician prescribing patterns. Unfortunately, Assumptions 4 and 5 only bound the treatment response over time between two conditional expectations or densities, respectively. We thus present algorithms that pinpoint the *exact* values for the CATE. See Appendix 7.4 for the derivation of the OCH$_d$ algorithm that pinpoints the CDTE.

### 4.1. CATE with Two Time Steps

We first consider the ideal scenario, where we have access to two time steps worth of observational data. The CATE is equivalent to the following under Assumption 4:

$$\begin{aligned}
&\mathbb{E}^r(Y_1(1)|\boldsymbol{X}) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X})\\
&\overset{4}{=} \Big[\mathbb{E}^o(Y_1|\boldsymbol{X},T=1)\mu_1 + \mathbb{E}^o(Y_0|\boldsymbol{X})(1-\mu_1)\Big] - \Big[\mathbb{E}^o(Y_1|\boldsymbol{X},T=0)\mu_0 + \mathbb{E}^o(Y_0|\boldsymbol{X})(1-\mu_0)\Big]\\
&= \psi_\mu(\boldsymbol{X})
\end{aligned}$$

for $\mu = (\mu_0, \mu_1) \in [0,1]^2$. The number above the equality sign references the assumption. Note that $\psi_\mu(\boldsymbol{X})$ applies to all of $\mathcal{S}_O$, since the three conditional expectations on the right hand side are derived from the observational distribution.

The quantities $\mu_0$ and $\mu_1$ however remain unknown. Fortunately, the CATE is equivalent to $\mathbb{E}^r(Y_1(1)|\boldsymbol{X},T=1,S=1) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X},T=0,S=1)$ in the RCT distribution under Assumptions 1 and 3 because $\mathbb{E}^r(Y_1(1)|\boldsymbol{X},T=1,S=1) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X},T=0,S=1) \overset{1}{=} \mathbb{E}^r(Y_1(1)|\boldsymbol{X},S=1) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X},S=1) \overset{3}{=} \mathbb{E}^r(Y_1(1)|\boldsymbol{X}) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X})$. We can therefore fit $\mu_0$ and $\mu_1$ on the area of overlap $\mathcal{S}_O \cap \mathcal{S}_R = \mathcal{S}_R$ by Assumption 2 using the trial data. We in particular minimize the distance to the CATE on $\mathcal{S}_R$ by solving:

$$\begin{aligned}
\mu^* = \arg\min_{\mu}\ & \mathbb{E}^r_{\boldsymbol{X}|S=1}\Big(g(\boldsymbol{X}) - \psi_\mu(\boldsymbol{X})\Big)^2\\
s.t.\ \ & 0 \leq \mu \leq 1,
\end{aligned} \tag{2}$$

where the outer expectation is taken over $\mathcal{S}_R$. We are now ready to state a main result:

**Theorem 1.** *$\psi_{\mu^*}(\boldsymbol{X})$ is equivalent to the CATE on $\mathcal{S}_O$ under Assumptions 1-4.*

*Proof.* The CATE is equivalent to $\psi_\mu(\boldsymbol{X})$ on $\mathcal{S}_O$ for some $\mu \in [0,1]^2$ by Assumption 4. The CATE is also equivalent to $\mathbb{E}^r(Y_1(1)|\boldsymbol{X},T=1,S=1) - \mathbb{E}^r(Y_1(0)|\boldsymbol{X},T=0,S=1)$ in the RCT distribution. The quantity $\psi_\mu(\boldsymbol{X})$ is unique on $\mathcal{S}_O$ for any $\mu \in \mathbb{R}^2$. The solution $\mu^*$ solving Expression (2) is unique on $\mathcal{S}_R$ because $\mathcal{S}_O \cap \mathcal{S}_R = \mathcal{S}_R$ by Assumption 2. The quantity $\psi_{\mu^*}(\boldsymbol{X})$ is therefore equivalent to the CATE on $\mathcal{S}_O$. □

We of course must estimate all necessary conditional expectations, in addition to $\mu_0$ and $\mu_1$, using the observational and trial data. Estimating the necessary conditional expectations and $\mu$ leads to the OCH$_2$ algorithm summarized in Algorithm 1. OCH$_2$ first estimates the CATE on $\mathcal{S}_R$ using the RCT data in Step 1. The algorithm then estimates each entry of $H(\boldsymbol{X}) = \{\mathbb{E}^o(Y_1|\boldsymbol{X}, T = 1), \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 0), \mathbb{E}^o(Y_0|\boldsymbol{X})\}$ using the observational data in Step 2. Next, OCH$_2$ approximates $\mu^*$ using the trial data in Step 3 by solving the empirical version of Expression (2) with the solutions of Steps 1 and 2:

$$\begin{aligned} \widehat{\mu} = \arg\min_{\mu} \frac{1}{2n} \sum_{i=1}^{2n} \Big( \overbrace{\widehat{g}(\boldsymbol{x}_i)}^{\text{Step 1}} - \overbrace{\widehat{\psi}_\mu(\boldsymbol{x}_i)}^{\text{Step 2}} \Big)^2 \\ s.t.\ \ 0 \leq \mu \leq 1, \end{aligned} \tag{3}$$

where $\widehat{\psi}_\mu(\boldsymbol{x}_i) = \Big[\widehat{\mathbb{E}}^o(Y_1|\boldsymbol{x}_i, T = 1)\mu_1 + \widehat{\mathbb{E}}^o(Y_0|\boldsymbol{x}_i)(1-\mu_1)\Big] - \Big[\widehat{\mathbb{E}}^o(Y_1|\boldsymbol{x}_i, T = 0)\mu_0 + \widehat{\mathbb{E}}^o(Y_0|\boldsymbol{x}_i)(1-\mu_0)\Big]$, and $n$ refers to the RCT sample size per treatment – assumed to be the same per treatment for notational convenience. The algorithm finally predicts $\widehat{\psi}_{\widehat{\mu}}(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the test set $\mathcal{T}$ each lying anywhere in $\mathcal{S}_O$.

**Algorithm 1:** Optimum in Convex Hulls with Two Time Steps (OCH$_2$)

```
  Input: trial data, observational data, test points $\mathcal{T}$
  Output: $\widehat{\psi}_{\widehat{\mu}}(\boldsymbol{X})$ on $\mathcal{T}$
1 Estimate $g(\boldsymbol{X})$ on $\mathcal{S}_R$ using the trial data
2 Estimate each entry of $H(\boldsymbol{X})$ on $\mathcal{S}_O$ using the observational data
3 Solve Expression (3) using $\widehat{g}(\boldsymbol{X})$ and $\widehat{H}(\boldsymbol{X})$ on the trial data
4 Predict $\widehat{\psi}_{\widehat{\mu}}(\boldsymbol{X})$ on $\mathcal{T}$
```

### 4.2. CATE with One Time Step

We unfortunately do not always have access to two time steps of observational data. Suppose however that the potential outcomes $Y_M(T)$ for $M = 0, 1$ and $T = 0, 1$ are appropriately normalized so that they are bounded below by zero; we can almost always satisfy this condition in clinical practice. We then have $\mathbb{E}(Y_0|\boldsymbol{X}) \geq 0$. If we take the worst case scenario $\mathbb{E}(Y_0|\boldsymbol{X}) = 0$, then Assumption 4 boils down to:

**Assumption 4′.** $\mathbb{E}^r(Y_1(T)|\boldsymbol{X}) = \mathbb{E}^o(Y_1|\boldsymbol{X}, T)\mu_T$ *where* $\mu_T \in [0, 1]$.

In other words, the above statement relaxes Assumption 4 from the convex hull of $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)$ and $\mathbb{E}^o(Y_0|\boldsymbol{X}) \geq 0$ to the wider convex hull of $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)$ and 0. Recovering the CATE then proceeds exactly as in Algorithm 1, but by setting $\widehat{\mathbb{E}}^o(Y_0|\boldsymbol{X})$ to zero. We refer to this variant as OCH$_1$ for one time step.

## 5. Experiments

We now investigate the accuracy of OCH using both synthetic and real data. Code is available at https://github.com/ericstrobl/OCH.

## 5.1. Algorithms

*State of the Art.* See Appendix 7.5 for a comprehensive discussion of related work. We compare $\text{OCH}_2$ against (1a) $\text{OCH}_1$ as well as six other algorithms representing the state of the art in CATE estimation under strict exclusion criteria: (2a) regression with RCT data only, (3a) regression with observational data only, (4a) OLT (Jackson et al., 2017), (5a) 2Step (Kallus et al., 2018), (6a) the conditional version of Difference in Differences (CDD) (Abadie, 2005), (7a) SDD (Strobl and Lasko, 2021). We compare $\text{OCH}_d$ for CDTE estimation with (1b) conditional density estimation with RCT data only and (2b) conditional density estimation with observational data only, since all other algorithms can only estimate the CATE. We will use the acronym RCT to refer to (2a) or (1b), and OBS to refer to (3a) or (2b), when it is clear that we mean the algorithms and not the datasets.

*Ablation Studies.* $\text{OCH}_1$ is an ablated version of $\text{OCH}_2$ obtained by removing the pre-treatment time step. We also compare the OCH variants for the CATE against (8a) $\text{UNC}_2$, or $\text{OCH}_2$ with the constraint in Expression (3) removed, and similarly (9a) $\text{UNC}_1$, or $\text{OCH}_1$ with the constraint removed. For the CDTE, we introduce (3b) $\text{UNC}_d$, or $\text{OCH}_d$ with the constraint in Expression (6) removed.

Note that the algorithms use different machine learning algorithms to estimate the required conditional expectations or densities out of box. We are however interested in isolating the performance of each algorithm independent of the chosen regressor or conditional density estimator. We therefore instantiate all algorithms with kernel ridge regression to estimate the required conditional expectations and the least squares probabilistic classifier (discrete outcome) or Dirac delta regression (continuous outcome) to estimate the required conditional densities in non-parametric form (Yamada et al., 2011; Strobl and Visweswaran, 2021). We equip both methods with the infinite knot spline kernel (Vapnik, 1998; Izmailov et al., 2013). We select the $\lambda$ hyperparameter for kernel ridge regression from the set {1E-8, 1E-7, . . . , 1E-1} and otherwise use default hyperparameters for the least squares probabilistic classifier and Dirac delta regression.

## 5.2. Synthetic Data

### 5.2.1. SIMULATION

We generate synthetic data using a mixture model. We sample the observational data i.i.d. from the following distribution:

$$Y_M(T) \sim \mathcal{N}(f_{MT}(Z), 0.1)$$

with $Z = \sum_{i=1}^{p} X_i$, each $X_i \sim \mathcal{U}(-1, 1)$ and the function $f_{MT}(Z)$ sampled uniformly from the set $\{Z, Z\Psi(Z), \exp(-Z^2), \tanh(Z)\}$ for $M = 0, 1$ and $T = 0, 1$. We sample the RCT data from:

$$Y_1(T) \sim \mu_T \mathcal{N}(f_{1T}(Z), 0.1) + (1 - \mu_T)\mathcal{N}_2(Z),$$

where $\mu_T \sim \mathcal{U}(0, 1)$ and $\mathcal{N}_2(Z) = \frac{1}{2}\mathcal{N}(f_{01}(Z), 0.1) + \frac{1}{2}\mathcal{N}(f_{00}(Z), 0.1)$. The density $p^r(Y_1(T)|\boldsymbol{X})$ is therefore a mixture of $p^o(Y_1|\boldsymbol{X}, T)$ and $p^o(Y_0|\boldsymbol{X}) = \frac{1}{2}p^o(Y_0|\boldsymbol{X}, T = 1) + \frac{1}{2}p^o(Y_0|\boldsymbol{X}, T = 0)$ satisfying Assumptions 4 and 5. See Appendix 7.7 for simulation results when the assumptions are violated.

We generate 1000 samples for the observational data split evenly between the two treatments and two time steps. We also generate 100 samples for the trial data split evenly between the two

| | 0% | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|
| OCH$_2$ | **0.0449** | **0.0503** | **0.0529** | **0.0550** | **0.0550** | **0.0609** |
| OCH$_1$ | 0.0555 | **0.0564** | **0.0564** | **0.0589** | **0.0578** | **0.0649** |
| UNC$_2$ | **0.0520** | 0.0653 | 0.0710 | 0.0819 | 0.0996 | 0.1225 |
| UNC$_1$ | 0.0606 | 0.0664 | 0.0695 | 0.0827 | 0.0962 | 0.1067 |
| SDD | 0.1266 | 0.1266 | 0.1364 | 0.1453 | 0.1569 | 0.1699 |
| 2Step | 0.2146 | 0.2136 | 0.2333 | 0.2634 | 0.2848 | 0.2879 |
| OBS | 0.2560 | 0.2492 | 0.2491 | 0.2570 | 0.2591 | 0.2631 |
| RCT | 0.1507 | 0.1949 | 0.2531 | 0.3331 | 0.3724 | 0.3615 |
| OLT | 0.2606 | 0.3226 | 0.4232 | 0.5204 | 0.5829 | 0.6799 |
| CDD | 0.6399 | 0.6476 | 0.6145 | 0.6267 | 0.6201 | 0.6514 |

(a)

| | 0% | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|
| OCH$_d$ | **0.1440** | **0.1376** | **0.1417** | **0.1401** | **0.1374** | **0.1461** |
| UNC$_d$ | **0.1448** | **0.1384** | **0.1442** | **0.1426** | **0.1410** | **0.1477** |
| OBS | 0.2304 | 0.2354 | 0.2482 | 0.2434 | 0.2340 | 0.2392 |
| RCT | 0.2566 | 0.2718 | 0.3128 | 0.3839 | 0.4529 | 0.4560 |

(b)

Table 1: Accuracy results for (a) the CATE and (b) the CDTE approximately sorted from best to worst. Lower is better. The OCH variants outperform the state of the art across all percentages of excluded subjects. Ablation studies reveal that the regression constraints $\mu_T \in [0,1]$ but not the pre-treatment data are necessary to achieve optimal performance with increasing exclusivity.

treatments. We impose strict inclusion criteria onto the trial data by excluding $r = 0, 25, 50, 75, 90$ or $95\%$ of patients by sampling $X_1$ according to $\mathcal{U}(-1 + 0.02r, 1)$; for example, excluding $50\%$ of patients is equivalent to sampling from $\mathcal{U}(0,1)$ on $50\%$ of the support of $\mathcal{U}(-1,1)$. We repeat the above procedure $500$ times for the CATE with the excluded percentages and $p = 1, 2, 6$ or $10$ variables in $\boldsymbol{X}$. We therefore generate a total of $500 \times 6 \times 4 = 12000$ independent datasets. We also repeat the above procedure $100$ times for the CDTE for a total of $100 \times 6 \times 4 = 2400$ datasets. We finally compare the algorithms by either computing the median of the mean squared error (MSE) to the ground truth CATE, or the median of the MISE to the ground truth CDTE; we use the median instead of the mean because the MSE and MISE histograms are skewed to the right for some algorithms.

### 5.2.2. PERFORMANCE

*Accuracy.* We summarize the results for the CATE in Table 1 (a) with algorithms roughly sorted from best to worst. Bolded values correspond to the best performance according to Mood's median test at a Bonferroni corrected $p$-value threshold of 0.05/9, since we ultimately compare each OCH variant against 9 other algorithms. When Assumption 4 holds, both OCH$_2$ and OCH$_1$ outperform all of their predecessors (2a-7a) across all percentages of excluded subjects (Table 1 (a)) and all numbers of variables in $\boldsymbol{X}$ (Table 2 (a) in Appendix 7.6). The algorithms even outperform RCT only with zero percent excluded subjects by taking advantage of the larger sample size of the observational dataset. Moreover, the constraints in OCH$_2$ and OCH$_1$ improve performance in most cases. Similarly, OCH$_2$ and OCH$_1$ outperform all other algorithms except their unconstrained variants when Assumption 4 is violated (Tables 3 (a) and 3 (b) in Appendix 7.7).

We summarize results for the CDTE in Table 1 (b). OCH$_d$ outperforms its competitors by a large margin regardless of whether Assumption 5 holds. The constraints however add little value when estimating the CDTE; OCH$_d$ and UNC$_d$ perform comparably across all proportions of excluded patients (Table 1 (b), and Table 4 (a) in Appendix 7.7) and across most variable numbers (Table 2 (b) in Appendix 7.6 and Table 4 (b) in Appendix 7.7). Densities must be non-negative and integrate to one, so constraining the mixing coefficients offers some but ultimately minimal additional benefit.

*Stability.* Consistently good performance, i.e. stability across datasets, is important for high stakes areas like medicine. When Assumption 4 holds, OCH prevents the median MSE from growing even with the vast majority of patients excluded, while the UNC variants do not (Figure 3 (a)). The deterioration in performance of RCT only is in fact much worse than even UNC$_2$ and UNC$_1$; the median MSE quickly increases with more stringent exclusion criteria (Figure 3 (b)).

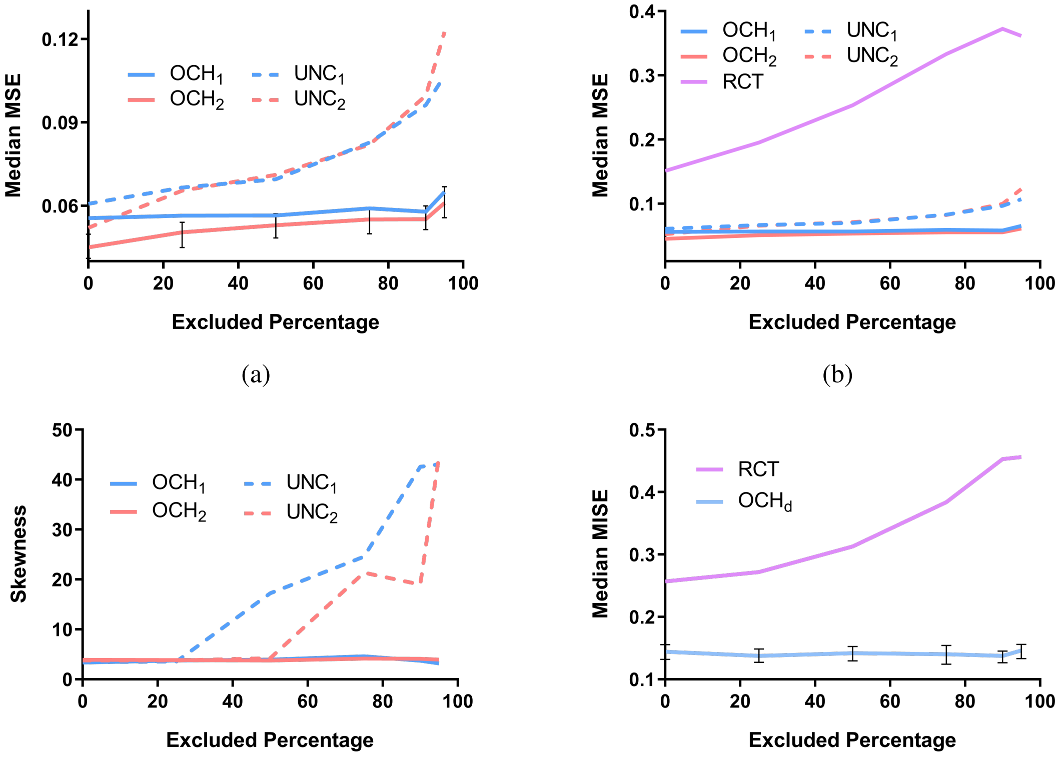
Closer inspection of the histograms show that the CATE OCH variants avoid catastrophic failures (very high MSE values) with higher percentages of excluded patients because the histogram of MSE values has minimal skewness to the right (to very large MSE values) (Figure 3 (c)). In contrast, skewness quickly increases for the UNC variants when the majority of patients cannot enter the RCT. We conclude that the constraints in Expression (3) are important for stability, particularly for very exclusive RCTs. Results when Assumption 4 is violated are qualitatively similar and presented in Appendix 7.7.

OCH$_d$ also prevents the MISE from growing as a larger proportion of patients are excluded from the RCT when Assumption 5 holds, unlike the RCT only algorithm; these results replicate those seen with the CATE (Figure 3 (d)). However, UNC$_d$ does not significantly increase skewness, again highlighting the minimal benefit of the constraints in the conditional density setting. Similar results apply when Assumption 5 is violated (Appendix 7.7).

### 5.3. Real Data

Evaluating the algorithms on real data is difficult because we rarely have access to the ground truth CATE or CDTE across the entire clinical population. Fortunately, investigators have conducted a handful of large, trans-institutional, multi-million dollar RCTs imposing few exclusion criteria. We use these RCTs to estimate the true CATE and CDTE. We then mimic more common exclusionary RCTs by imposing additional exclusion criteria. We finally generate observational data by asking a physician to remove patients who fail to match common prescribing patterns from the original RCTs. We present results for one real dataset here and refer the reader to Appendix 7.8 for a second real dataset.

Many randomized trials exclude patients who use other medications or illegal substances. Nevertheless, these patients are often the sickest. We therefore evaluated the algorithms on how well they generalize to the broader population when trained on an RCT excluding this sub-group and a confounded observational dataset.



(a) (b)

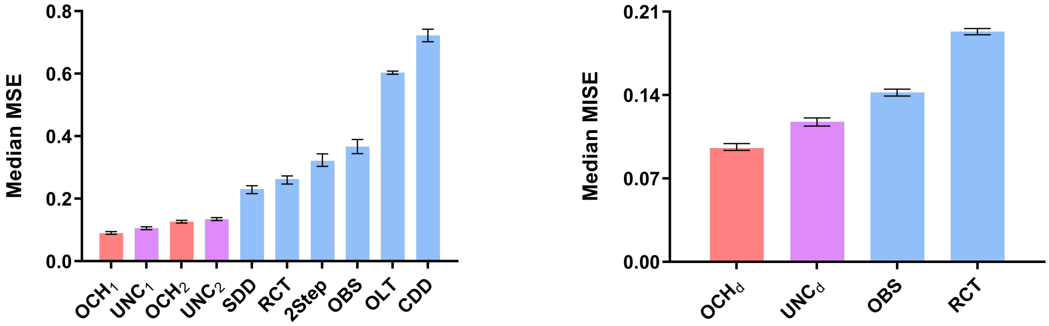
(c) (d)

Figure 3: Stability results when Assumptions 4 and 5 hold. Lower and flatter is better. Error bars denote 95% confidence intervals for $\text{OCH}_2$ or $\text{OCH}_d$ but may not be visible. (a) The constraints in $\text{OCH}_2$ and $\text{OCH}_1$ prevent the MSE from growing with a higher percentage of excluded patients. (b) RCT offers terrible stability. (c) The constraints in $\text{OCH}_2$ and $\text{OCH}_1$ in particular prevent catastrophic failures by constraining skewness to the right (very high MSE values) in very exclusive trials. (d) $\text{OCH}_d$ also controls the MISE in the density setting, whereas RCT does not.

We obtained data from the Sequenced Treatment Alternatives to Relieve Depression trial (STAR*D), a large inclusive RCT designed to assess the sequential effects of anti-depressants and cognitive therapy on patients with major depressive disorder (Rush et al., 2006; Warden et al., 2007; Trivedi et al., 2006). The investigators assessed treatment response using QIDS-SR, a self-reported measure of depressive symptoms. STAR*D ultimately included four levels of sequential treatment assignment.

We analyzed data from the second level because it had a large sample size and tested the effects of buproprion ($T = 1$) versus venlafaxine and sertraline ($T = 0$). Buproprion is known to have a unique effect on a symptom of depression called hypersomnia, or excessive sleepiness (Papakostas et al., 2006). We examined the hypersomnia sub-score of QIDS-SR at week 6 in order to give sufficient time for the treatments to elicit differential effect. We used the other sub-scores in QIDS-SR related to sleep as predictors in $\boldsymbol{X}$, including sleep onset insomnia, mid-nocturnal insomnia and early morning insomnia. We treated this dataset containing 388 samples as the comprehensive RCT.

We generated observational data by imposing confounding on the comprehensive RCT. Physicians often prefer buproprion for patients with major depression who experience hypersomnia. We



(a)

(b)

Figure 4: Real data results. We plot (a) median MSE values and (b) median MISE values for STAR*D. Error bars again denote 95% confidence intervals. The OCH algorithms in red usually achieve the best performance.

therefore removed patients receiving $T = 1$ with no hypersomnia, or a hypersomnia QIDS-SR score of zero, but kept all patients receiving $T = 0$. This process ultimately excluded 19.3% of the original 388 patients.

We next generated exclusive RCT data by imposing additional exclusion criteria. We in particular performed a literature search and identified (1) current psychotropic use and (2) substance use as the two most common exclusion criteria in clinical trials of major depression not already implemented in STAR*D (Blanco et al., 2017). We therefore excluded patients meeting at least one of those two criteria. This process eliminated 39.8% of the original 388 patients.

We finally ran all of the algorithms on 2000 bootstrapped draws of the derived observational and exclusive RCT datasets. We quantify accuracy using either the MSE to the ground truth CATE, or the MISE to the ground truth CDTE estimated using all of the original 388 patients. For the CATE, both OCH$_2$ and OCH$_1$ outperform their predecessors (Figure 4 (a)). The ablated variants UNC$_2$ and UNC$_1$ also perform well, but not as well as their constrained counterparts. The results for the CDTE are similar; OCH$_d$ performs the best, followed by UNC$_d$ (Figure 4 (b)). We conclude that all OCH algorithms perform well in estimating the CATE or CDTE even when including patients who use other psychotropics or substances (or both).

## 6. Conclusion

Physicians identify seemingly effective treatments by observing patient outcomes over time; they then readily administer those treatments to certain patients. We used this observation to propose a new approach to cross-design synthesis, where we bound the unconfounded treatment response between the confounded treatment response seen before and after treatment assignment. This implies that the treatment effect must lie in the convex hull of two sets of conditional expectations or densities. We exploited the assumptions in three variants of the OCH algorithm which all analyze RCT and observational data simultaneously in order recover either the CATE or the CDTE over the entire population. Experimental results highlighted the superior performance of OCH compared to its predecessors. We conclude that OCH offers a promising new approach to generalizing randomized trials.

## References

Lithium for suicidal behavior in mood disorders. In *ClinicalTrials.gov Identifier: NCT01928446*. National Library of Medicine, 2021. URL https://clinicaltrials.gov/ct2/show/NCT01928446.

Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelth AAAI National Conference on Artificial Intelligence*, page 230–237. AAAI Press, 1994.

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

Scott R Beach, Federico Gomez-Bernal, Jeff C Huffman, and Gregory L Fricchione. Alternative treatment strategies for catatonia: A systematic review. *General Hospital Psychiatry*, 48:1–19, 2017.

Carlos Blanco, Nicolas Hoertel, Silvia Franco, Mark Olfson, Jian-Ping He, Saioa López, Ana González-Pinto, Frédéric Limosin, and Kathleen R Merikangas. Generalizability of clinical trial results for adolescent major depressive disorder. *Pediatrics*, 140(6), 2017.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Technical report, National Bureau of Economic Research, 1993.

Judith Droitcour, George Silberman, and Eleanor Chelimsky. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care*, 9(3):440–449, 1993.

Frederick K Goodwin, Bruce Fireman, Gregory E Simon, Enid M Hunkeler, Janelle Lee, and Dennis Revicki. Suicide risk in bipolar disorder during treatment with lithium and divalproex. *JAMA*, 290(11):1467–1473, 2003.

Joseph F Hayes, Alexandra Pitman, Louise Marston, Kate Walters, John R Geddes, Michael King, and David PJ Osborn. Self-harm, unintentional injury, and suicide in bipolar disorder during maintenance mood stabilizer treatment: a uk population-based electronic health records study. *JAMA Psychiatry*, 73(6):630–637, 2016.

Rauf Izmailov, Vladimir Vapnik, and Akshay Vashist. Multidimensional splines with infinite number of knots as svm kernels. In *The 2013 International Joint Conference on Neural Networks*, pages 1–7. IEEE, 2013.

Christopher Jackson, John Stevens, Shijie Ren, Nick Latimer, Laura Bojke, Andrea Manca, and Linda Sharples. Extrapolating survival from randomized trials using external data: a review of methods. *Medical Decision Making*, 37(4):377–390, 2017.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 10888–10897. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/566f0ea4f6c2e947f36795c8f58ba901-Paper.pdf.

Erik Lauterbach, Werner Felber, B Müller-Oerlinghausen, B Ahrens, Thomas Bronisch, T Meyer, B Kilb, Ute Lewitzka, Barbara Hawellek, A Quante, et al. Adjunctive lithium treatment in the prevention of suicidal behaviour in depressive disorders: a randomised, placebo-controlled, 1-year trial. *Acta Psychiatrica Scandinavica*, 118(6):469–479, 2008.

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

Charles F Manski. Anatomy of the selection problem. *Journal of Human Resources*, pages 343–360, 1989.

Joseph P McEvoy, Jonathan M Meyer, Donald C Goff, Henry A Nasrallah, Sonia M Davis, Lisa Sullivan, Herbert Y Meltzer, John Hsiao, T Scott Stroup, and Jeffrey A Lieberman. Prevalence of the metabolic syndrome in patients with schizophrenia: baseline results from the clinical antipsychotic trials of intervention effectiveness (catie) schizophrenia trial and comparison with national estimates from nhanes iii. *Schizophrenia Research*, 80(1):19–32, 2005.

Maria A Oquendo, Hanga C Galfalvy, Dianne Currier, Michael F Grunebaum, Leo Sher, Gregory M Sullivan, Ainsley K Burke, Jill Harkavy-Friedman, M Elizabeth Sublette, Ramin V Parsey, et al. Treatment of suicide attempters with bipolar disorder: a randomized clinical trial comparing lithium and valproate in the prevention of suicidal behavior. *American journal of psychiatry*, 168 (10):1050–1056, 2011.

George I Papakostas, David J Nutt, Lindsay A Hallett, Vivian L Tucker, Alok Krishen, and Maurizio Fava. Resolution of sleepiness and fatigue in major depressive disorder: a comparison of bupropion and the selective serotonin reuptake inhibitors. *Biological Psychiatry*, 60(12):1350–1355, 2006.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

A John Rush, Madhukar H Trivedi, Stephen R Wisniewski, Andrew A Nierenberg, Jonathan W Stewart, Diane Warden, George Niederehe, Michael E Thase, Philip W Lavori, Barry D Lebowitz, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a star*d report. *American Journal of Psychiatry*, 163(11):1905–1917, 2006.

Linmarie Sikich, Jean A Frazier, Jon McClellan, Robert L Findling, Benedetto Vitiello, Louise Ritz, Denisse Ambler, Madeline Puglia, Ann E Maloney, Emily Michael, et al. Double-blind comparison of first-and second-generation antipsychotics in early-onset schizophrenia and schizo-affective disorder: findings from the treatment of early-onset schizophrenia spectrum disorders (teoss) study. *American Journal of Psychiatry*, 165(11):1420–1431, 2008.

Eric V Strobl. Improved causal discovery from longitudinal data using a mixture of dags. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, pages 100–133. PMLR, 2019.

Eric V Strobl. Causal discovery with a mixture of dags. *Machine Learning*, pages 1–25, 2022.

Eric V. Strobl and Thomas A. Lasko. Synthesized difference in differences. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384506. doi: 10.1145/3459930.3469528. URL https://doi.org/10.1145/3459930.3469528.

Eric V Strobl and Shyam Visweswaran. Dirac delta regression: Conditional density estimation with clinical trials. In *The KDD'21 Workshop on Causal Discovery*, pages 78–125. PMLR, 2021.

T Scott Stroup, Jeffrey A Lieberman, Joseph P McEvoy, Sonia M Davis, Marvin S Swartz, Richard SE Keefe, Alexander L Miller, Robert A Rosenheck, John K Hsiao, CATIE Investigators, et al. Results of phase 3 of the catie schizophrenia trial. *Schizophrenia Research*, 107(1): 1–12, 2009.

Elizabeth A Stuart, Catherine P Bradshaw, and Philip J Leaf. Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16(3):475–485, 2015.

Madhukar H Trivedi, A John Rush, Stephen R Wisniewski, Andrew A Nierenberg, Diane Warden, Louise Ritz, Grayson Norquist, Robert H Howland, Barry Lebowitz, Patrick J McGrath, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in star*d: implications for clinical practice. *American journal of Psychiatry*, 163(1):28–40, 2006.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Diane Warden, A John Rush, Madhukar H Trivedi, Maurizio Fava, and Stephen R Wisniewski. The star*d project results: a comprehensive review of findings. *Current psychiatry reports*, 9(6): 449–459, 2007.

Makoto Yamada, Masashi Sugiyama, Gordon Wichern, and Jaak Simm. Improving the accuracy of least-squares probabilistic classifiers. *IEICE Transactions on Information and Systems*, 94(6): 1337–1340, 2011.

## 7. Appendix

### 7.1. Graphical Characterization

We provide a graphical characterization of Assumptions 1-3 using Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013). We prefer SWIGs over twin networks (Balke and Pearl,

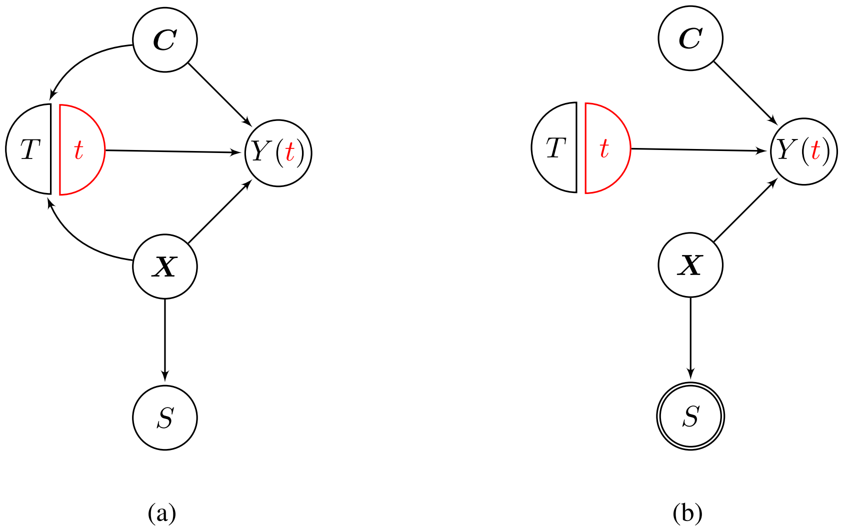


Figure 5: SWIGs for the observational distribution in (a) and the RCT distribution in (b).

1994), since SWIGs encode all conditional independence relations – among the variables present in the SWIG – that hold for all distributions over counterfactuals. Twin networks lack this completeness property and were designed to model additional cross-world independencies, which we are not interested in in this paper.

We first require some definitions. A *directed graph* $\mathbb{G}$ is a graph over a set of vertices $\boldsymbol{Z}$ with at most one directed edge between any two vertices. A *directed path* from $A$ to $B$ is a sequence of directed edges from $A$ to $B$. A *cycle* occurs when there exists a directed path from $A$ to $B$ and $B \rightarrow A$. A *directed acyclic graph* (DAG) is a directed graph without cycles. $A$ is an *ancestor* of $B$ if there exists a directed path from $A$ to $B$. A *collider* refers to the triple $A \rightarrow B \leftarrow C$. Two vertices $A$ and $B$ are *d-connected* given $\boldsymbol{W} \subseteq (\boldsymbol{Z} \setminus \{A, B\})$ in a DAG, if there exists a path between $A$ and $B$ such that any collider on the path is an ancestor of $\boldsymbol{W}$ and no non-collider on the path is in $\boldsymbol{W}$. Otherwise, $A$ and $B$ are *d-separated* given $\boldsymbol{W}$.

We summarize the causal relations using the SWIG of the observational distribution in Figure 5 (a); a SWIG is a DAG where we split the treatment vertex $T$ into $T$ and $t$ and replace the outcome $Y$ with the counterfactual variable $Y(t)$. The vertex $\boldsymbol{C}$ denotes a set of potentially unobserved confounders. We obtain the SWIG of the RCT distribution in Figure 5 (b) by removing the directed edges into $T$. The d-separation relation between $Y(t)$ and $T$ given $(\boldsymbol{X}, S)$ in Figure 5 (b), and likewise the d-separation relation between $\boldsymbol{X}$ and $T$ given $S$ imply Assumption 1. Assumption 3 corresponds to the d-separation relation between $Y(t)$ and $S$ given $\boldsymbol{X}$. The second assumption holds from the fact that $S = 1$ in the RCT distribution but can take on any value in the observational distribution.

## 7.2. CATE versus CDTE

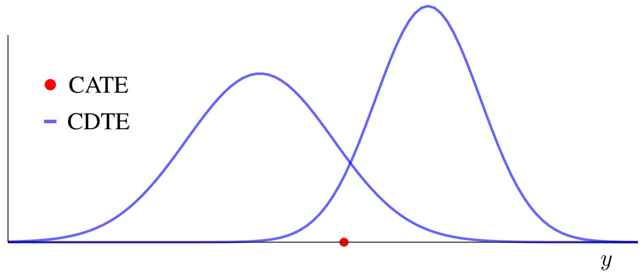


Figure 6: The CATE only provides a point estimate for each patient, whereas the CDTE summarizes the probabilities across *all* possible outcome values. We do not require Gaussianity for the CDTE.

## 7.3. Counter-Example

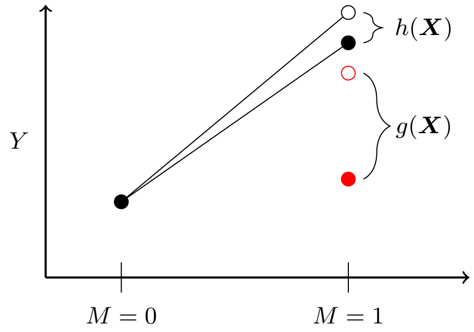


Figure 7: Example of a situation where $g(\boldsymbol{X})$ is larger than $h(\boldsymbol{X}) = \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 1) - \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 0)$ but Assumption 4 holds. The two black lines correspond to the change in expected outcomes of the two treatments in the observational distribution, while the two red dots correspond to the expected outcomes of the two treatments in the RCT distribution at $M = 1$.

## 7.4. Conditional Densities of Treatment Effect

### 7.4.1. CDTE WITH TWO TIME STEPS

The CATE only provides a point estimate of the differential effect of treatment. As mentioned previously, we would like to visualize the probabilities associated with all possible outcome values using the CDTE. Assumption 5 helps us recover the CDTE using two time steps. The assumption requires $p^o(Y_1|\boldsymbol{X}, T)$ and $p^o(Y_0|\boldsymbol{X})$, which we can obtain from the observational distribution. The mixing proportion $\mu_T \in [0, 1]$ remains unspecified, but we have $p^r(Y_1(T)|\boldsymbol{X}, T, S = 1) = p^r(Y_1(T)|\boldsymbol{X}, S = 1) = p^r(Y_1(T)|\boldsymbol{X})$ in the RCT distribution by Assumptions 1 and 3, respectively. We can therefore solve for the optimal $\mu_T^*$ by minimizing the following mean integrated squared error (MISE) quantifying the distance between the mixture density $p^o(Y_1|\boldsymbol{X}, T)\mu_T+$

$p^o(Y_0|\boldsymbol{X})(1-\mu_T)$ and the desired density $p^r(Y_1(T)|\boldsymbol{X})$:

$$\begin{aligned}\mu_T^* = \arg\min_{\mu_T} \mathbb{E}^r_{\boldsymbol{X}|S=1}\Big[\int \big([p^o(Y_1 = y|\boldsymbol{X},T)\mu_T + p^o(Y_0 = y|\boldsymbol{X})(1-\mu_T)] &\\ - p^r(Y_1(T) = y|\boldsymbol{X})\big)^2\, dy\Big]&\\ s.t.\;\; 0 \leq \mu_T \leq 1,&\end{aligned} \tag{4}$$

where the expectation is taken with respect to the RCT distribution, or on $\mathcal{S}_R$. We can solve Expression (4) without access to $p^r(Y_1(T)|\boldsymbol{X})$ because the objective function is proportional to:

$$\begin{aligned}&\frac{1}{2}\mathbb{E}^r_{\boldsymbol{X}|S=1}\Big[\int p^o(Y_1|\boldsymbol{X},T)^2 + p^o(Y_0|\boldsymbol{X})^2 - 2p^o(Y_1|\boldsymbol{X},T)p^o(Y_0|\boldsymbol{X})\; dy\Big]\mu_T^2 - \\ &\Big[\underbrace{\mathbb{E}^r_{Y_1(T),\boldsymbol{X}|S=1}\Big(p^o(Y_1|\boldsymbol{X},T) - p^o(Y_0|\boldsymbol{X})\Big)}_{\text{Expectation w.r.t. } \mathbb{P}^r(Y_1(T)|\boldsymbol{X})\mathbb{P}^r(\boldsymbol{X}|S=1)} - \mathbb{E}^r_{\boldsymbol{X}|S=1}\Big(\int p^o(Y_1|\boldsymbol{X},T)p^o(Y_0|\boldsymbol{X}) - p^o(Y_0|\boldsymbol{X})^2\Big)\Big]\mu_T,\end{aligned} \tag{5}$$

obtained by expanding the square and dropping constants that do not depend on $\mu_T$. This new form does not require access to $p^r(Y_1(T)|\boldsymbol{X})$, but only to the expectation with respect to $\mathbb{P}^r(Y_1(T)|\boldsymbol{X}, S = 1)\mathbb{P}^r(\boldsymbol{X}|S = 1) = \mathbb{P}^r(Y_1(T)|\boldsymbol{X})\mathbb{P}^r(\boldsymbol{X}|S = 1)$ as indicated by the underbrace per Assumption 3. We therefore obtain the same $\mu_T^*$ by replacing the objective function in Expression (4) with the one above.

We have the result below to seal the strategy by following a similar argument as Theorem 1 but extended to conditional densities:

**Theorem 2.** *The quantity $p^o(Y_1|\boldsymbol{X},T)\mu_T^* + p^o(Y_0|\boldsymbol{X})(1-\mu_T^*)$ is equivalent to $p^r(Y_1(T)|\boldsymbol{X})$ on $\mathcal{S}_O$ under Assumptions 1-3 and 5.*

*Proof.* $p^r(Y_1(T)|\boldsymbol{X})$ is equivalent to $p^o(Y_1|\boldsymbol{X},T)\mu_T + p^o(Y_0|\boldsymbol{X})(1-\mu_T)$ for some $\mu_T \in [0,1]$ by Assumption 5. We also know that $p^r(Y_1(T)|\boldsymbol{X},T,S = 1)$ is equivalent to $p^r(Y_1(T)|\boldsymbol{X})$ in the RCT distribution under Assumptions 1 and 3: $p^r(Y_1(T)|\boldsymbol{X},T,S = 1) \overset{1}{=} p^r(Y_1(T)|\boldsymbol{X},S = 1) \overset{3}{=} p^r(Y_1(T)|\boldsymbol{X})$. The quantity $p^o(Y_1|\boldsymbol{X},T)\mu_T + p^o(Y_0|\boldsymbol{X})(1-\mu_T)$ is unique on $\mathcal{S}_O$ for any $\mu_T \in \mathbb{R}$. The solution $\mu_T^*$ to Expression (4) is unique on $\mathcal{S}_R$ because $\mathcal{S}_O \cap \mathcal{S}_R = \mathcal{S}_R$ by Assumption 2. The quantity $p^o(Y_1|\boldsymbol{X},T)\mu_T^* + p^o(Y_0|\boldsymbol{X})(1-\mu_T^*)$ is therefore equivalent to $p^r(Y_1(T)|\boldsymbol{X})$ on $\mathcal{S}_O$. □

We present the corresponding algorithm called Optimum in Convex Hulls for Densities (OCH$_d$) for the finite sample setting in Algorithm 2. The method is similar to Algorithm 1 with some important differences. First, OCH$_d$ does not directly approximate the CDTE using the RCT. The algorithm instead approximates $p^o(Y_0|\boldsymbol{X})$ in Step 1 and $p^o(Y_1|\boldsymbol{X},T)$ in Step 3 using the observational data. OCH$_d$ then obtains the empirical estimate of $\mu_T^*$ in Step 4 by solving the following quadratic program – equivalent to the empirical version of Expression (5), obtained by replacing expectations with means and densities with their estimates:

$$\begin{aligned}\widehat{\mu}_T = \arg\min_{\mu_T}\;\; &\frac{1}{2}H\mu_T^2 - d\mu_T\\ s.t.\;\; &0 \leq \mu_T \leq 1,\end{aligned} \tag{6}$$

where:

$$H = \frac{1}{2n}\sum_{i=1}^{2n}\Big[\int \widehat{p}^o(Y_1 = y|\boldsymbol{x}_i, T)^2 + \widehat{p}^o(Y_0 = y|\boldsymbol{x}_i)^2 - 2\widehat{p}^o(Y_1 = y|\boldsymbol{x}_i, T)\widehat{p}^o(Y_0 = y|\boldsymbol{x}_i)\, dy\Big]$$

$$d = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{p}^o(Y_1 = y_{1i}(T)|\boldsymbol{x}_i) - \widehat{p}^o(Y_0 = y_{1i}(T)|\boldsymbol{x}_i)\Big)$$

$$- \frac{1}{2n}\sum_{i=1}^{2n}\Big(\int \widehat{p}^o(Y_1 = y|\boldsymbol{x}_i, T)\widehat{p}^o(Y_0 = y|\boldsymbol{x}_i) - \widehat{p}^o(Y_0 = y|\boldsymbol{x}_i)^2\, dy\Big).$$

Finally, the algorithm predicts $\widehat{p}^o(Y_1|\boldsymbol{x}, T)\widehat{\mu}_T + \widehat{p}^o(Y_0|\boldsymbol{x})(1 - \widehat{\mu}_T)$ for all $\boldsymbol{x}$ in the test set $\mathcal{T}$ each lying anywhere in $\mathcal{S}_O$.

**Algorithm 2:** Optimum in Convex Hulls for Densities (OCH$_d$)

```
Input: trial data, observational data, test points $\mathcal{T}$
Output: $\widehat{p}^o(Y_1|\boldsymbol{X}, t)\widehat{\mu}_t + \widehat{p}^o(Y_0|\boldsymbol{X})(1 - \widehat{\mu}_t)$ on $\mathcal{T}$ for $t \in \{0, 1\}$
Estimate $p^o(Y_0|\boldsymbol{X})$
for $t \in \{0, 1\}$ do
    Estimate $p^o(Y_1|\boldsymbol{X}, t)$ on $\mathcal{S}_O$ using the observational data
    Solve Expression (6) using $\widehat{p}^o(Y_1|\boldsymbol{X}, t)$ and $\widehat{p}^o(Y_0|\boldsymbol{X})$ on the trial data
    Predict $\widehat{p}^o(Y_1|\boldsymbol{X}, t)\widehat{\mu}_t + \widehat{p}^o(Y_0|\boldsymbol{X})(1 - \widehat{\mu}_t)$ on $\mathcal{T}$
end
```

#### 7.4.2. CDTE WITH ONE TIME STEP

Estimating the CDTE with one time step is unfortunately not a straightforward modification of the CDTE with two time steps as with the CATE. The problem stems from the indeterminancy of a "worst case scenario" for $p^o(Y_0|\boldsymbol{X})$. Simply choosing the Dirac delta function at zero does not work because, if $p^r(Y_1(T)|\boldsymbol{X})$ is contained in the convex hull of $p^o(Y_1|\boldsymbol{X}, T)$ and $p^o(Y_0|\boldsymbol{X})$, then $p^r(Y_1(T)|\boldsymbol{X})$ is *not* necessarily contained in the convex hull of $p^o(Y_1|\boldsymbol{X}, T)$ and the Dirac delta. In contrast, if $\mathbb{E}^r(Y_1(T)|\boldsymbol{X})$ is contained in the convex hull of $\mathbb{E}^o(Y_1|\boldsymbol{X}, T)$ and $\mathbb{E}^o(Y_0|\boldsymbol{X})$, then $\mathbb{E}^r(Y_1(T)|\boldsymbol{X})$ is clearly contained in the convex hull of $E^o(Y_1|\boldsymbol{X}, T)$ and 0 because $\mathbb{E}^o(Y_0|\boldsymbol{X}) \geq 0$ with suitable normalization. The logic with conditional expectations therefore does not carry over to conditional densities. Estimating the CDTE with one time step likely requires a non-linear generalization of the convex hull, so we leave it open to future work.

### 7.5. Related Work

The variants of OCH fall into a category of methods that accomplish *cross-design synthesis*, *transportability*, *data fusion* or *generalizability*; these terms refer to the act of combining trial and observational data (and potentially other dataset types) in order to the eliminate the weaknesses of each (Droitcour et al., 1993; Stuart et al., 2015; Bareinboim and Pearl, 2016). Note that a plethora of algorithms, such as inverse probability weighted estimators, exist when $\mathcal{S}_O \subseteq \mathcal{S}_R$, but relatively few methods investigate the more realistic situation when $\mathcal{S}_R \subset \mathcal{S}_O$ with *strict* exclusion criteria in randomized trials (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004). The earliest methods in the latter category assume unconfoundedness and simply estimate the CATE using the observational data (Rosenbaum and Rubin, 1983; Pearl, 2009). The weakness of these methods of course lies in the unconfoundedness assumption, which we can neither guarantee nor verify in practice.

Later algorithms proposed to modify the observational estimate $f(\boldsymbol{X}) = \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 1) - \mathbb{E}^o(Y_1|\boldsymbol{X}, T = 0)$ using a linear transformation. The earliest algorithm in this category, which we call Outer Linear Transform (OLT), proposed:

$$g(\boldsymbol{X}) = f(\boldsymbol{X})\alpha + \beta,$$

where $\alpha$ and $\beta$ are fit using linear regression with trial data (Jackson et al., 2017). OLT has the desirable property of preserving the shape of $f(\boldsymbol{X})$, but it offers limited flexibility in adjusting $f(\boldsymbol{X})$. A subsequent method called 2Step proposed to modify $f(\boldsymbol{X})$ using a linear combination of the predictors (Kallus et al., 2018):

$$g(\boldsymbol{X}) = f(\boldsymbol{X}) + \boldsymbol{X}\delta + \beta.$$

It remains unclear however *why* $f(\boldsymbol{X})$ should be linearly related to $g(\boldsymbol{X})$ via $\boldsymbol{X}$. Subsequent authors therefore proposed a more principled approach to choosing the basis functions in an algorithm called Synthesized Difference in Differences (SDD) (Strobl and Lasko, 2021). SDD linearly combines four conditional expectations as follows:

$$\begin{aligned} g(\boldsymbol{X}) =& [\mathbb{E}^o(Y_1|\boldsymbol{X}, T = 1) - \mathbb{E}^o(Y_0|\boldsymbol{X}, T = 1)\alpha_1] \\ &- [\mathbb{E}^o(Y_1|\boldsymbol{X}, T = 0)\alpha_2 - \mathbb{E}^o(Y_0|\boldsymbol{X}, T = 0)\alpha_3], \end{aligned}$$

The authors showed that this transformation relaxes the parallel slopes assumption used in the conditional Difference in Differences algorithm (Abadie, 2005). SDD therefore carries theoretical justification, but it performs unstably in practice; the algorithm requires a substantial amount of regularization in order to consistently estimate the CATE with a high degree of accuracy, calling into question whether the underlying assumption actually holds in practice.

The OCH variants synthesize RCT and observational data by exploiting fundamentally different assumptions than those adopted by prior methods – Assumptions 4 or 5. The OCH algorithms in particular assume that confounding may exacerbate the magnitude of the treatment response over time, but it preserves the direction; this weakens the unconfoundedness assumption which requires that both the magnitude and direction are preserved once we condition on $\boldsymbol{X}$. OCH thus provides a theoretical explanation as to *why* the adopted basis functions should be linearly adjusted in order to recover the CATE. OCH also introduces regularization naturally into Expressions (2) and (4) via the convex hull. Finally, the algorithms extend the CATE and the CDTE to $\mathcal{S}_O$, whereas prior methods only extend the CATE. OCH therefore offers a superior approach to cross-design synthesis compared to its predecessors.

More generally, OCH capitalizes on a mixture of distributions which has already been exploited in causal graph learning with large gains in performance relative to methods that assume a single distribution (Strobl, 2019, 2022). OCH was originally inspired by this mixture framework, even though the variants tackle a different problem. If accounting for mixtures improves performance with causal graph learning, then it should also improve performance with treatment effect estimation.

### 7.6. Additional Experimental Results when Assumptions Hold

| | 1 | 3 | 6 | 10 |
|---|---|---|---|---|
| $\text{OCH}_2$ | **0.0186** | **0.0353** | **0.0703** | **0.1170** |
| $\text{OCH}_1$ | **0.0214** | **0.0439** | **0.0824** | **0.1302** |
| $\text{UNC}_2$ | 0.0637 | 0.0465 | 0.0821 | 0.1318 |
| $\text{UNC}_1$ | 0.0517 | 0.0485 | 0.0914 | 0.1458 |
| SDD | 0.0627 | 0.0962 | 0.1666 | 0.2336 |
| 2Step | 0.0742 | 0.1767 | 0.3320 | 0.4987 |
| RCT | 0.1525 | 0.1735 | 0.2500 | 0.4009 |
| OBS | 0.0628 | 0.1545 | 0.3155 | 0.5248 |
| OLT | 0.3314 | 0.3922 | 0.4908 | 0.6732 |
| CDD | 0.2551 | 0.4681 | 0.8057 | 1.2309 |

(a)

| | 1 | 3 | 6 | 10 |
|---|---|---|---|---|
| $\text{OCH}_d$ | **0.0272** | **0.1157** | **0.1906** | **0.2425** |
| $\text{UNC}_d$ | 0.0319 | **0.1166** | **0.1914** | **0.2428** |
| OBS | 0.1042 | 0.2024 | 0.2693 | 0.3072 |
| RCT | 0.2895 | 0.3191 | 0.3719 | 0.3703 |

(b)

Table 2: Accuracy results for (a) the CATE and (b) the CDTE across all numbers of variables in $\boldsymbol{X}$ when Assumptions 4 and 5 hold.

### 7.7. Experimental Results when Assumptions are Violated

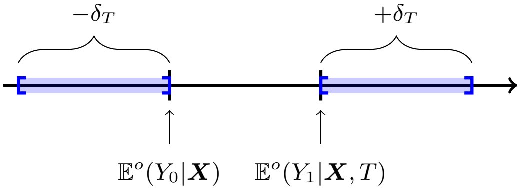
We repeat the simulation experiments but consider violations of Assumptions 4 and 5. Let $\delta_T = E(Y_1(T)|T, \boldsymbol{X}) - E(Y_0|\boldsymbol{X})$. We sample the RCT data with probability 0.5 from:

$$Y_1(T) \sim \mu_T \mathcal{N}(f_{1T}(Z) + \delta_T, 0.1) + (1 - \mu_T)\mathcal{N}(f_{1T}(Z), 0.1),$$

and with the other probability 0.5 from:

$$Y_1(T) \sim \mu_T \mathcal{N}_2(Z) + (1 - \mu_T)\mathcal{N}_2'(Z),$$

where $\mathcal{N}_2'(Z) = \frac{1}{2}\mathcal{N}(f_{01}(Z), 0.1) + \frac{1}{2}\mathcal{N}(f_{00}(Z), 0.1) - \delta_T$. Pictorially, these situations correspond to sampling from outside the assumed convex hull:


$-\delta_T$
$+\delta_T$
$\mathbb{E}^o(Y_0|\boldsymbol{X})$
$\mathbb{E}^o(Y_1|\boldsymbol{X}, T)$

We summarize accuracy results in Tables 3 and 4, and stability results in Figure 8.

| | 0% | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|
| UNC$_1$ | **0.1580** | **0.1698** | **0.2162** | **0.2314** | **0.2729** | 0.3226 |
| UNC$_2$ | **0.1471** | **0.1667** | **0.2259** | **0.2598** | 0.3248 | 0.3540 |
| OCH$_2$ | 0.1892 | **0.1878** | **0.2106** | **0.2418** | **0.2730** | **0.2818** |
| OCH$_1$ | 0.2017 | 0.2000 | **0.2008** | **0.2280** | **0.2377** | **0.2553** |
| SDD | 0.2643 | 0.2659 | 0.3061 | 0.3222 | 0.3693 | 0.3898 |
| 2Step | 0.3407 | 0.3417 | 0.3677 | 0.3494 | 0.3705 | 0.3820 |
| OBS | 0.3707 | 0.3882 | 0.3793 | 0.3659 | 0.4010 | 0.3733 |
| RCT | 0.2928 | 0.3309 | 0.4457 | 0.6141 | 0.7680 | 0.7425 |
| CDD | 0.8735 | 0.8356 | 0.8895 | 0.8491 | 0.8895 | 0.8983 |
| OLT | 0.5674 | 0.6011 | 0.7160 | 0.8168 | 0.9671 | 1.0516 |

(a)

| | 1 | 3 | 6 | 10 |
|---|---|---|---|---|
| UNC$_2$ | 0.1329 | **0.1279** | **0.2927** | **0.4822** |
| UNC$_1$ | **0.0999** | **0.1265** | **0.3036** | **0.4857** |
| OCH$_2$ | **0.0779** | **0.1453** | **0.3313** | **0.5392** |
| OCH$_1$ | **0.0763** | 0.1488 | **0.3375** | **0.5361** |
| SDD | 0.1106 | 0.2017 | 0.4388 | 0.6770 |
| 2Step | 0.0695 | 0.2288 | 0.5481 | 0.9180 |
| RCT | 0.1935 | 0.3347 | 0.5576 | 0.8427 |
| OBS | 0.0800 | 0.2361 | 0.6107 | 1.0483 |
| OLT | 0.5561 | 0.6117 | 0.9586 | 1.2268 |
| CDD | 0.3054 | 0.6240 | 1.1115 | 1.7799 |

(b)

Table 3: Accuracy results for the CATE when Assumption 4 is violated. Sub-figures (a) and (b) correspond to different percentages of excluded subjects and numbers of variables in $\boldsymbol{X}$, respectively. The two algorithms OCH$_2$ and OCH$_1$ are only outperformed by their unconstrained variants when Assumption 4 fails to hold.

### 7.8. Additional Real Data Results

Many randomized trials exclude children, but children get sick too. We therefore evaluated the algorithms on how well they generalize to children when trained on an RCT only recruiting adults and a confounded observational dataset.

We in particular obtained data from two clinical trials investigating the effects of anti-psychotics on schizophrenia spectrum disorders. The CATIE trial recruited 530 adults who were at least 18 years old, while the TEOSS trial recruited 62 children up to 19 years old (McEvoy et al., 2005; Sikich et al., 2008; Stroup et al., 2009).

Clinicians prefer olanzapine ($T = 1$) over risperidone ($T = 0$) for excited patients – defined as hyperactivity, heightened responsivity, hyper vigilance or excessive mood lability – because olanzapine is more sedating. We therefore set the excitement subscore of the PANSS scale, a quantitative measure of schizophrenia symptoms, at week 4 as the outcome. We then predicted differences in

| | 0% | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|
| OCH$_d$ | **0.2008** | **0.2084** | **0.1964** | **0.2030** | **0.1973** | **0.1940** |
| UNC$_d$ | **0.2047** | **0.2134** | **0.2063** | **0.2106** | **0.2056** | **0.2052** |
| OBS | 0.2341 | 0.2405 | 0.2460 | 0.2379 | 0.2366 | 0.2372 |
| RCT | 0.3210 | 0.3242 | 0.3625 | 0.4064 | 0.4636 | 0.4890 |

(a)

| | 1 | 3 | 6 | 10 |
|---|---|---|---|---|
| OCH$_d$ | **0.0687** | **0.1620** | **0.2361** | **0.3016** |
| UNC$_d$ | 0.0864 | **0.1649** | **0.2391** | **0.3068** |
| OBS | 0.0992 | 0.1993 | 0.2714 | **0.3104** |
| RCT | 0.3255 | 0.3755 | 0.3869 | 0.4118 |

(b)

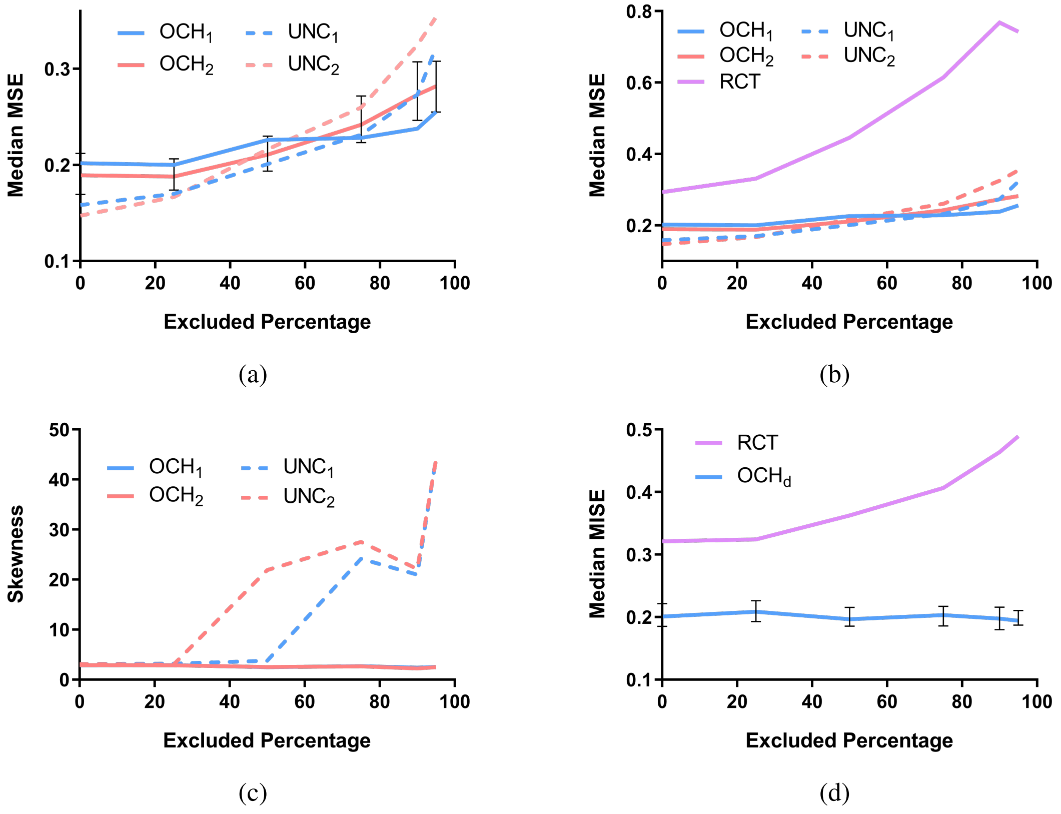
Table 4: Accuracy results for the CDTE when Assumption 5 does not hold. OCH$_d$ outperforms RCT and observational data only for all (a) percentages of excluded subjects, and (b) numbers of variables in $\boldsymbol{X}$.

excitement using age and the PANSS hostility sub-score as predictors in $\boldsymbol{X}$ because adults who are hostile are more dangerous than weaker children. The comprehensive RCT corresponds to the combined CATIE and TEOSS dataset consisting of $530 + 62 = 592$ patients.

We generated the observational data by first combining CATIE and TEOSS. We then excluded patients assigned to $T = 1$ who were not excited (excitement subscore less than or equal to two – the formal cutoff for questionable pathology), and patients assigned to $T = 0$ who were excited (subscore greater than two). This mimics real world prescribing patterns where physicians prescribe olanzapine to excited patients.

We used the 530 patients in the CATIE trial as the exclusive RCT data. Since we include age as a predictor, the goal is to generalize to children. We therefore tested the algorithms on their ability to accurately predict the CATE or CDTE on children using this adult RCT and the confounded observational dataset.

We report the results over 2000 bootstrapped draws in Figures 9 (a) and 9 (b). OCH$_2$ and OCH$_1$ again outperformed their predecessors and their unconstrained variants (Figure 9 (a)). Similar results held with OCH$_d$, although OCH$_d$ did not outperform UNC$_d$ in this case similar to the synthetic data results (Figure 9 (b)). The performance improvements were much larger in this dataset compared to STAR*D. Finally notice that the RCT only algorithm performs terribly (median MSE 33.2 and median MISE 0.31) because non-linear regressors or conditional density estimators cannot consistently extrapolate well from adults to children, or to unseen regions on the support $S_O \setminus S_R$.



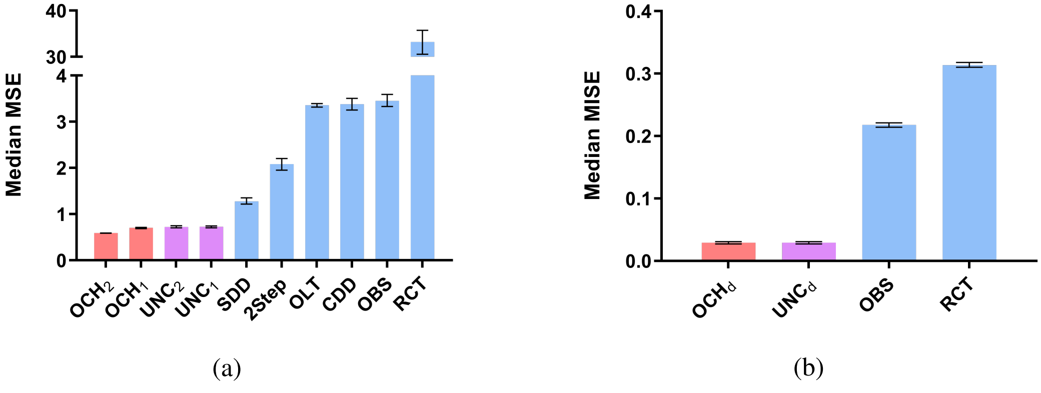
(a)

(b)

(c)

(d)

Figure 8: Stability results when Assumptions 4 and 5 are violated. Results are qualitatively similar to those in Figure 3.



(a)

(b)

Figure 9: The (a) median MSE values and (b) median MISE values for CATIE and TEOSS. The OCH algorithms again achieve the best performance.