How and Why to Use Experimental Data to Evaluate Methods for Observational Causal Inference

Amanda Gentzel [1 2] **Purva Pruthi** [1] **David Jensen** [1]

Abstract

Methods that infer causal dependence from observational data are central to many areas of science, including medicine, economics, and the social sciences. A variety of theoretical properties of these methods have been proven, but *empirical* evaluation remains a challenge, largely due to the lack of observational data sets for which treatment effect is known. We describe and analyze *observational sampling from randomized controlled trials* (OSRCT), a method for evaluating causal inference methods using data from randomized controlled trials (RCTs). This method can be used to create constructed observational data sets with corresponding unbiased estimates of treatment effect, substantially increasing the number of data sets available for empirical evaluation of causal inference methods. We show that, in expectation, OSRCT creates data sets that are equivalent to those produced by randomly sampling from empirical data sets in which all potential outcomes are available. We then perform a large-scale evaluation of seven causal inference methods over 37 data sets, drawn from RCTs, as well as simulators, real-world computational systems, and observational data sets augmented with a synthetic response variable. We find notable performance differences when comparing across data from different sources, demonstrating the importance of using data from a variety of sources when evaluating any causal inference method.

1. Introduction

Researchers in machine learning and statistics have become increasingly interested in methods that estimate causal effects from observational data. Such interest is understandable, given the centrality of causal questions in fields such as medicine, economics, sociology, and political science (Morgan & Winship, 2015). Causal inference has also emerged as an important class of methods for improving the explainability and fairness of machine learning systems, since causal models can explicitly represent the underlying mechanisms of systems and their likely behavior under counterfactual conditions (Kusner et al., 2017; Pearl, 2019).

However, evaluating causal inference methods is far more challenging than evaluating purely associational methods. Both types of methods can be analyzed theoretically. However, *empirical* analysis—long a driver of research progress in machine learning and statistics—has been increasingly recognized as vital for research progress in causal inference (e.g., Dorie et al., 2019; Gentzel et al., 2019), and empirical evaluation is substantially more challenging to perform in the case of causal inference. Many associational models (e.g., classifiers and conditional probability estimators) can be evaluated using cross-validation or held-out test sets. However, causal inference aims to estimate the value or distribution of an outcome variable *under intervention*, and evaluating such estimates requires an alternative route to estimating the effects of such interventions.

Most available data sets are either experimental (which can yield unbiased estimates of treatment effect) or observational (for which treatment effect is unknown). Since most causal inference methods are designed to infer causal dependence from observational data, accurate evaluation requires both observational data and corresponding unbiased estimates of treatment effect. Several recent efforts have attempted to address this problem (e.g., Dorie et al., 2019; Gentzel et al., 2019; Tu et al., 2019; Shimoni et al., 2018), most of which collect or modify data specifically for the purpose of evaluation. Some approaches induce dependence between variables in specially constructed or selected data, while others re-purpose a simulator to produce data for evaluation. These approaches are promising and beneficial to the community, but creating individual, specialized new data sets is difficult and time-consuming, limiting the number of data sets available. To this point, most data sets that are easy to collect (such as synthetic data and simulators) are generally unrealistic, while data sets with a high degree of

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, United States [2]Leidos, Reston, Virginia, United States. Correspondence to: Amanda Gentzel <Amanda.M.Gentzel@leidos.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

realism require significant effort to obtain.

We argue for exploiting a largely untapped source of data for evaluating causal inference methods: randomized controlled trials (RCTs). RCTs are designed and conducted for the express purpose of providing unbiased estimates of treatment effect. Many RCT data sets are publicly available, and that number is only increasing, allowing for the collection of a large number of realistic data sets. Previous work has described how to non-randomly sample a specialized type of experimental data (one in which all potential outcomes are observed) to create *constructed observational data sets*.[1] Surprisingly, this basic approach can be modified to produce constructed observational data from RCTs as well.

The core idea of non-random sampling of data from RCTs is not original to this paper. This basic approach has been used sporadically for at least a decade to evaluate algorithms for observational causal inference (Hill, 2011). However, these uses have typically been described only in passing and have often been limited to a single data set. This basic approach has also been used more widely to evaluate algorithms for learning policies for contextual bandits (Li et al., 2011). However, this work is almost unknown within the community of researchers studying observational causal inference. The properties and wide utility of this approach are sufficiently unknown that a more general analysis and discussion is warranted. Section 3 provides additional discussion of related work.

We use data from multiple RCTs and also from a wide variety of other data sources to perform a large-scale evaluation of several causal inference methods. These data sources can be broadly categorized as RCTs, simulators (Guillaume & Rougemont, 2006; Tu et al., 2019; Miller et al., 2020), real-world computational systems (Gentzel et al., 2019) and observational data sets augmented with a synthetic response variable (Dorie et al., 2019; Shimoni et al., 2018). This type of large-scale evaluation from a diverse set of data sources provides us with an unprecedented opportunity to analyze systematic differences in the performance of the methods not only due to characteristics of the algorithms but also due to differences in data source.

Specifically, we: (1) Propose that a known, but sporadically used, method for inducing confounding bias in RCT data become part of the standard evaluation suite for causal inference methods; (2) Prove that this approach is equivalent, in expectation, to the data generating process assumed by the potential-outcomes framework, a longstanding theoretical framework for causal inference; (3) Demonstrate the feasibility of this approach by applying multiple causal inference methods to observational data constructed from RCTs; and (4) Perform a large-scale evaluation of seven causal inference methods on 37 data sets drawn from multiple empirical data sources.[2]

2. Creating Observational Data from Randomized Controlled Trials

Consider a data generating process that produces a binary treatment $T \in \{0, 1\}$, outcome Y, and multiple covariates $C = \{C_1, C_2, ...C_k\}$, each of which may be causal for outcome.[3] We define $Y_i(t)$ to be the outcome for unit i under treatment t, referred to as a *potential outcome*. For each unit i, both treatment values $T_i = 0$ and $T_i = 1$ are set by intervention and both potential outcomes $Y_i(1)$ and $Y_i(0)$ are measured. We refer to this type of data, where all potential outcomes are observed, as *all potential outcomes* (APO) data, denoted D_{APO}. Note that, due to the use of explicit interventions, such a data generating process produces *experimental*, rather than observational, data.

Recently, some researchers (Louizos et al., 2017; Gentzel et al., 2019) have proposed sampling from APO data to produce constructed observational data. Such data sets are produced by probabilistically sampling a treatment value (and its corresponding outcome value) for every unit based on the values of one or more covariates. We refer to $C^b \subseteq C$ as the *biasing covariates* and selected treatment random variable as T^s, thus $P(T^s|C^b) := f(C^b)$. This procedure, shown in Algorithm 1, induces causal dependence between C^b and T^s, creating a confounder when C^b also causes Y. We refer to such a data generating process as *observational sampling from all potential outcomes* (OSAPO) and denote a given data set generated in this way as D_{OSAPO}. OSAPO is the data generating process assumed under the potential outcomes framework (Rubin, 2005).

Data sets produced by OSAPO are extremely useful for evaluating causal inference methods. Causal inference methods can estimate treatment effect in D_{OSAPO}, and these estimates can be compared to estimates derived from D_{APO}. Furthermore, the process of inducing bias by sub-sampling allows for a degree of control that can be exploited to evaluate a method's resilience to confounding, by systematically varying the strength and form of dependence and whether variables in C^b are observed. However, very few experimental data sets exist that record all potential outcomes for every

[1] The term "constructed observational data" denotes empirical data to which additional properties common in observational data (e.g., confounding) have been synthetically introduced. This term is distinct from *constructed observational studies*, which are studies that collect and compare both experimental and observational data from the same domain (see Section 3).

[2] Pointers to the data sets used in this paper, and R code to perform observational sampling, are available at https://github.com/KDL-umass/papers/tree/main/causal-eval-rct-icml-2021.

[3] For ease of exposition, we describe the approach using binary treatment, but the approach is more general.

unit, severely limiting the applicability of this approach.

2.1. Observational Sampling of RCTs

Now consider a slightly different data generating process, in which treatment is randomly assigned and only one potential outcome is measured for each unit i, producing either $Y_i(1)$ or $Y_i(0)$, but not both. This is the data generating process typically implemented by RCTs, in which every unit is randomly assigned a single treatment value, and the outcome for that treatment is measured. Vast numbers of RCTs are conducted each year, and data sets from many of them are available publicly. In addition, growing efforts toward open science are continually increasing the number of publicly available RCT data sets.

This raises an intriguing research question: *Can RCTs be sub-sampled to produce constructed observational data sets with the same properties as those produced by APO sampling?*

We describe one such sampling procedure in Algorithm 2 — *observational sampling from randomized controlled trials* (OSRCT)—which produces a data sample denoted D_{OSRCT}. As in APO sampling, covariates C^b bias the selection of a single treatment value, t_s, for every unit i. If unit i actually received the selected treatment t_s, we add i to D_{OSRCT}. Otherwise, that unit is ignored. As we show below in Theorem 2, when treatment is binary and the treatment and control groups are equal in size, the resulting constructed observational data set is, in expectation, half the size of the original, regardless of the form of the biasing. As discussed in Section 2.2, a causal inference method can then be applied to this data, and the results can be compared to the unbiased effect estimate from the original RCT data. This basic approach is shown in Figure 2.

An RCT can be thought of as a data set where one potential outcome for every unit is missing at random. Since OSRCT uses the biasing covariates to select treatment, and treatment was assigned randomly, the sub-sampling process only changes the dependence between the biasing covariates and treatment. This is the same as in APO sampling. The probability of a given unit-treatment pair being included in the sub-sample is proportional in APO and RCT sampling. That is, D_{OSRCT} is equivalent to a random sample of D_{OSAPO}. We demonstrate this empirically, using an APO data set converted into an RCT data set, and provide a proof, in the supplementary material.

Theorem 1. *For RCT data set D_{RCT}, APO data set D_{APO}, and binary treatment $T \in \{0,1\}$ with $P(T=1) = P(T=0) = 0.5$ in D_{RCT}, and units i, $P_{D_{OSRCT}}(T_i = t) = 0.5 * P_{D_{OSAPO}}(T_i = t)$, for all units i and treatment values t.*

Note that while Theorem 1 assumes equal probability of treatment and control, the approach generally applies even when $P(T=1) \neq 0.5$. In this case, instead of sub-sampling D_{OSAPO} by a factor of 0.5, the scaling factor is selected based on the treatment value. Since treatment is based solely on the value of the biasing covariates, this is equivalent to modifying the form of the biasing function.

One potential disadvantage of this approach is that sub-sampling to induce bias necessarily reduces the size of the resulting sample. Somewhat surprisingly, however, the degree of this reduction does not depend on the intensity of the biasing.

Theorem 2. *For binary treatment $T \in \{0,1\}$ and RCT data set D_{RCT}, if either $P(T=1) = P(T=0) = 0.5$, or $E[P(T^s = 1|C^b)] = 0.5$, then $E[|D_{OSRCT}|] = 0.5|D_{RCT}|$.*

A proof is provided in the Supplementary Material.

One downside of OSRCT is that it reduces the sample size available for causal inference. One alternative is to *reweight* units, rather than sub-sample them, according to $P(T_i^s = t_i|C_i^b)$. This has the benefit of basing causal estimates on every unit in the RCT rather than only those in the sub-sample. This approach requires that the causal inference method under evaluation accepts unit-level weights. However, when this is an option, using weighting rather than sub-sampling can be a useful alternative when RCTs with small sample size [4]. We compared the estimates obtained by weighting to the estimates obtained by sub-sampling for causal inference methods that accept unit-level weights and found the results to be similar. Comparison results are provided in the supplementary material.

2.2. What Can OSRCT Evaluate?

The constructed observational data created by OSRCT has a substantial benefit over purely observational data: Unbiased estimates of causal effect can be obtained from the original RCT data, which can be compared to effect estimates from causal inference methods. A well-designed RCT enables the unbiased estimation of the sample average treatment effect (ATE) as $E[y_i(1)] - E[y_i(0)] = E[y_i|t_i = 1] - E[y_i|t_i = 0]$, where t_i denotes the actual treatment received by unit i. This estimate can be compared to estimates made by causal inference methods applied to the constructed observational data.

Unlike APO data, RCT data only contains one treatment-outcome pair for every unit, limiting both the available effect estimates and how these data sets can be used. RCTs measure the effect of a single randomized intervention $do(T_i = t_i)$ for every unit in the data set. Thus, we cannot estimate individual treatment effect (ITE) from RCT data, a

[4] Thanks to an anonymous reviewer for the recommendation of this approach

Algorithm 1 Observational sampling from all potential outcomes (OSAPO)

Input: APO data set D_{APO}, biasing covariates C^b
Output: Biased data set D_{OSAPO}
for all units $i \in D$ **do**
  $p \leftarrow f(C_i^b)$
  $t_s \leftarrow Bernoulli(p)$
  $o \leftarrow$ row in D_{APO} corresponding to (i, t_s)
  $D_{OSAPO} \leftarrow D_{OSAPO} \cup o$
end for
Return D_{OSAPO}

Algorithm 2 Observational sampling from randomized controlled trials (OSRCT)

Input: RCT data set D_{RCT}, biasing covariates C^b
Output: Biased data set D_{OSRCT}
for all unit $i \in D$ **do**
  $p \leftarrow f(C_i^b)$
  $t_s \leftarrow Bernoulli(p)$
  if $(i, t_s) \in D$ **then**
    $o \leftarrow$ row in D_{RCT} corresponding to (i, t_s)
    $D_{OSRCT} \leftarrow D_{OSRCT} \cup o$
  end if
end for
Return D_{OSRCT}

Figure 1. Two procedures for sampling constructed observational data sets from experimental data. *Left*: From all potential outcomes (APO) data. *Right*: From randomized controlled trial (RCT) data. For some function $f : \mathcal{D}(C^b) \rightarrow \{x \in \mathbb{R} : 0 < x < 1\}$

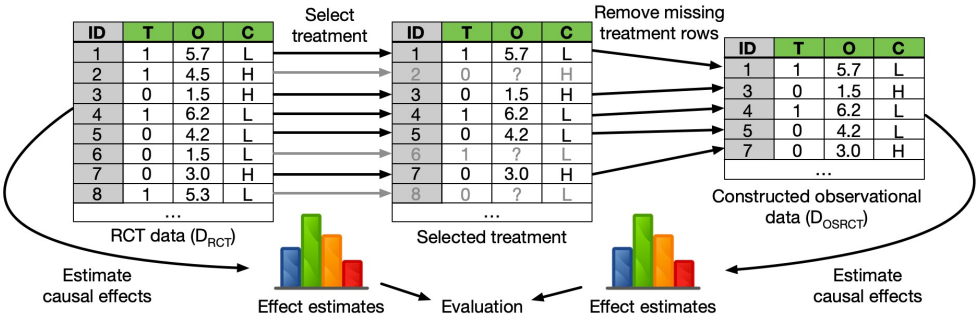

Figure 2. The process of creating observational-style data from a randomized controlled trial.

measurement that is available when using APO data.

However, while RCTs do not measure individual-level treatment effects, they do measure individual-level outcomes under intervention. Thus, OSRCT data *can* be used to evaluate a method's ability to estimate unit-level outcome under intervention. Any causal inference method that can estimate $E[Y|do(T = t)]$ can be evaluated by comparing those estimates against measurements in the RCT data. In addition, the data discarded during biased sub-sampling in OSRCT can be treated as a held out test set, where the dependence between the biasing covariates and treatment is the complement of that in D_{OSRCT}. We call the data discarded during OSRCT the "complementary sample". Because we know the functional form of the dependence between the biasing covariates and treatment, we can weight the data points in the complementary sample according to their probability of being included in the accepted sample. In aggregate, this type of weighting allows the complementary sample to approximate the distribution of the training data, and thus be used for testing. This is equivalent to inverse propensity score weighting (Rosenbaum & Rubin, 1983). The supplementary material contains more details on using the complementary sample for evaluation, including a proof.

2.3. Assumptions, Limitations, and Opportunities

The validity of evaluation with OSRCT depends on several standard assumptions about the validity of the original RCT. Specifically, it assumes that treatment assignment is randomized and that all sampled units complete the study (no "drop-out"). Intriguingly, one standard assumption—that intent to treat does not differ from actual treatment—is not necessary. Even if this assumption is violated, the estimated treatment effect will correspond to the effect of intending to treat, and this estimand can still be used to evaluate the effectiveness of methods for observational causal inference.

Evaluation with OSRCT has some limitations. OSRCT can induce dependence between any covariate and treatment, but not between any covariate and outcome. In addition, while the original RCT data can yield an unbiased estimate of the effect of treatment on outcome, it cannot produce such estimates for any other pair of variables.

Constructing observational data also provides some unique opportunities. OSRCT produces data with non-random treatment assignment, and allows for variation in the level and form of that non-randomness. Additional features of observational studies can also be simulated, such as measurement error, selection bias, positivity violations, and hidden confounding (by hiding one of the biasing covariates). While some of these may further reduce the sample size of the constructed observational data due to additional sub-sampling, this process of constructing observational data can allow for the evaluation of a causal inference method's robustness to many features of real-world data.

One potential concern about the use of RCTs is their realism. While RCTs are empirical, the situations in which RCTs are conducted may differ from many observational settings. For example, RCTs are performed in settings where treatment can be randomized, where relevant covariates can be collected, where outcome can be easily measured, and, in many cases, where there is an assumption that a causal effect is likely.

For the purpose of evaluation, though, the important question is not whether the causal effect estimates from the RCT data are representative of the population of interest, but whether the performance of causal inference methods on samples from the RCT data are representative of their performance on the population of interest. While this is not guaranteed, this is a far weaker assumption than requiring that the RCT population match the desired observational population.

3. Related Work

OSRCT has been used for at least a decade in two different settings. The first is the most similar to what we describe in this paper, in which OSRCT has been applied sporadically to evaluate methods for causal inference from observational data. An early use is Hill (2011), in which the author non-randomly samples data from an RCT—the Infant Health and Development Program (IHDP)—to produce a single observational-style data set for evaluating Bayesian additive regression trees (BART). The resulting constructed observational data set has subsequently been reused by others to evaluate a variety of methods for observational causal inference (e.g., Shalit et al., 2017; Atan et al., 2018; Yao et al., 2018; Saito & Yasui, 2020). Less frequently, researchers have applied OSRCT or closely related approaches to other data sets drawn from RCTs or natural experiments (e.g., Arceneaux et al., 2006; Kallus et al., 2018b; Kallus & Zhou, 2018; Witty et al., 2020; Zhang & Bareinboim, 2021).

Despite these sporadic uses, observational sampling from RCTs is not widely known or used within the causal inference community. For example, two recent papers that systematically review existing evaluations of causal inference methods—Dorie et al. (2019) and Gentzel et al. (2019)—do not even mention this approach, despite the fact that it overcomes many of the most serious threats to validity for evaluation studies (e.g., reproducibility, realistic data distributions and complexity of treatment effects, multiple possible levels of confounding). It has not been explicitly formalized nor have its advantages been clearly described. As a result, it is rarely used and it has never been systematically compared to alternative approaches to evaluation.

The second setting in which OSRCT has been applied is offline evaluation of contextual bandit policies (Li et al., 2011). Specifically, Li et al. show how to evaluate a (non-random) contextual bandit policy by sampling from the data produced by a randomized policy.This method is widely employed to evaluate methods in fields such as computational advertising and recommender systems (e.g., Tang et al., 2013; 2014; Zeng et al., 2016), and it has been extended with approaches such as bootstrapping (Mary et al., 2014).

Our use of OSRCT in this paper exploits the same idea but in a subtly different setting. In our setting, we have no interest in estimating the effect of a contextual policy that is known to the agent. Instead, our goal is to determine how well a given method estimates the average or conditional treatment effect (which, in contextual bandits, would be formulated as the reward difference between two specific policies), even though the algorithm only has access to the actions and outcomes of a single unknown and non-randomized policy.

In addition to these two settings that have directly used OSRCT, other prior work has explicitly focused on empirical evaluation methods for observational causal inference. The ideal method would score highly on at least three characteristics: *data availability* (many data sets with the required characteristics can be easily obtained to avoid overgeneralizing from a small sample); *internal validity* (differences between estimated treatment effect and the standard can only be attributed to bias in the estimator rather than biases due to how data sets were created); and *external validity* (the performance of the estimator will generalize well to other settings). Of three broad classes of prior work, each suffers from deficiencies and none clearly dominate.

Some prior work uses *observational data sets with known treatment effect*. One approach gathers observational data about phenomena that are so well-understood that the causal effect is obvious (e.g., Mooij et al., 2016), but such situations are rare, limiting data availability. Another approach uses data from matched pairs of observational and experimental studies (e.g., Dixit et al., 2016; Sachs et al., 2005; Jaciw, 2016). Such data sets appear to represent a nearly ideal scenario for evaluating methods for inferring causal effect from observational data, but pairs of directly comparable observational and experimental studies have low data

availability and using paired studies with different settings or variable definitions can greatly reduce internal validity. Some "constructed observational studies" intentionally create matched pairs of experimental and observational data sets (e.g., LaLonde, 1986; Hill et al., 2004; Shadish et al., 2008), but these studies also have low data availability.

Another class of prior work generates observational data from *synthetic or highly controlled causal systems* (e.g., Tu et al., 2019; Gentzel et al., 2019; Louizos et al., 2017; Kallus et al., 2018a; Athey et al., 2021). In this way, the treatment effect is either directly known or can be easily derived from experimentation. Observational data is typically obtained via some biased sampling of the experimental data, often a variety of APO sampling. In the case of entirely synthetic data, data availability and internal validity are both high, but external validity is low, and such studies are often criticized as little more than demonstrations. External validity typically increases somewhat for highly controlled causal systems, but data availability drops significantly.

The final and newest class of existing work augments *an existing observational study with a synthetic outcome*, replacing the original outcome measurement with the result of a synthetic function (e.g., Dorie et al., 2019; Shimoni et al., 2018). Given the synthetic nature of the outcome function, the causal effect is known. This class of approach has relatively high data availability, and it trades some loss of external validity (because real outcome measurements are replaced with synthetic ones) to gain internal validity (because the true treatment effect is known). Note particularly that *both* the treatment effect *and* the confounding are synthetic, because the function that determines the synthetic outcome determines how both the treatment and potential confounders affect the value of outcome.

The approach proposed here—OSRCT—augments, rather than replaces, existing approaches. It occupies a unique position because it simultaneously has fairly high data availability, internal validity, and external validity. OSRCT's data availability is relatively high because it can use data from any moderately large RCT. Only synthetic data generators and synthetic outcome approaches have higher data availability, but both suffer in terms of external validity. OSRCT's internal validity is relatively high because there exist many well-designed RCTs. Using synthetic data generators or highly controlled causal systems will typically produce somewhat higher internal validity, as will observational data with synthetic outcomes, but this is done at the cost of external validity or data availability. Finally, OSRCT's external validity is relatively high because the distributions of all variables and the outcome function occur naturally, while only the confounding is synthetic. Only observational studies with known treatment effect have higher external validity, and these suffer from severe limitations on data availability.

4. Are RCT Data Sets Available?

OSRCT has the benefit of leveraging existing empirical data rather than requiring the creation of new data sets specifically for evaluating causal inference methods, but it does require that data from RCTs be available and generally accessible to causality researchers. Fortunately, this is increasingly the case. While many repositories that host RCTs are restricted for reasons of privacy and security, many other repositories allow access with only minimal restrictions. In some cases, access requires only registering with the repository and agreeing not to re-distribute the data or attempt to de-anonymize it. As long as these data sharing agreements are adhered to, such data can be easily acquired by causality researchers. An even larger set of repositories restricts access but will make data available upon reasonable request. Additional information about some such repositories can be found in the supplementary material.

In addition, funding agencies and journals are increasingly requiring that researchers make anonymized individual patient data available upon reasonable request (Godlee & Groves, 2012; Ohmann et al., 2017). For example, the United States' National Institutes of Health (NIH) recently requested public feedback on a proposed data sharing policy with the aim of improving data management and the sharing of data created by NIH-funded projects (Request for Public Comment on DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance, 2019). There is also increasing awareness and discussion in the medical community about the importance of sharing individual patient data, to allow for greater transparency and re-analysis (Drazen, 2015; Kuntz et al., 2019; Banzi et al., 2019; Suvarna, 2015). All of this suggests increasing availability of individual patient data from randomized controlled trials.

5. Evaluation

To assess how well RCT data sets work for evaluation, we performed a large-scale evaluation using RCT data, as well as data from three additional sources: computational systems, synthetic-response data, and simulators.

5.1. Data

Computational Systems. We use the computational system data sets provided by Gentzel et al. (2019). In these data sets, all potential outcomes are observed, so we refer to these as APO data sets. These data sets are collected from three computational systems: queries executed by a Postgres database, HTTP requests executed by web servers on the open internet, and programs compiled under the Java Development Kit. For each data set, we selected a single treatment-outcome pair and a biasing covariate, matching the setup in Gentzel et al. (2019).

Synthetic-Response Data. Many data sets for evaluation were created for the ACIC Competition (Dorie et al., 2019) and the IBM Causal Inference Benchmarking Framework (Shimoni et al., 2018). These data sets were created using a set of real-world covariates and then simulating both a treatment and an outcome. As in APO and RCT data, treatment is selected synthetically based on values of one or more covariates. However, the outcome (sometimes referred to as the 'response surface') is also generated synthetically. Both the ACIC Competition and the IBM Causal Inference Benchmarking Framework created a large number of data sets, with varying treatment and outcome functions. We selected five data sets from each, for a total of ten data sets, to use for our evaluation.

Simulators. We used data sets from three simulators of varying complexity: a simulator of neuropathic pain (Tu et al., 2019); Nemo (Guillaume & Rougemont, 2006), a simulator of population dynamics; and three simple simulators from the WhyNot Python package (Miller et al., 2020). For both Nemo and the neuropathic pain simulator, we chose three distinct treatment-outcome pairs, generating three data sets for each. For the WhyNot simulators, we chose three separate simulators and generated a single data set from each, resulting in nine total data sets from simulators.

Randomized Controlled Trials. We selected data sets from six repositories: Dryad (2020), the Yale Institution for Social and Policy Studies Repository (2020), the NIH National Institute on Drug Abuse Data Share Website (2020), the University of Michigan's ICPSR repository (2020), the UK Data Service (2020), and the Knowledge Network for Biocomplexity (2020). These repositories were selected because they contained RCT data, were reasonably well-documented, and had a simple data access process. None of these repositories house RCT data exclusively, so some search and filtering was necessary to identify relevant data sets. We also used a stand-alone RCT from a study of artificial cultural markets (Salganik et al., 2006).

These four sources of data represent a range of realism. Of the simulators, the WhyNot simulators are the most simplistic and, by design, are not intended to be realistic representations of the world. The Nemo and neuropathic pain simulators are a bit more realistic, since they were designed by experts within their respective communities with the intent of accurately modeling their respective systems. The synthetic-response data sets are even more realistic, because the covariate distribution is from an empirical source. However, both the treatment and outcome are synthetic. The APO and RCT data sets have realistic covariate distributions and response functions, with synthetic treatment functions. While the APO data sets are limited to a narrow set of domains, the RCT data sets come from a wide range of studies in different fields. This spectrum allows us to assess how algorithm performance differs under different levels of realism. Further details about all of the data sets are included in the supplementary material.

5.2. Algorithms

Due to the nature of the ground truth in the selected data sets (treatment effect of a single treatment on a single outcome), we focused our evaluation on causal inference methods that estimate average treatment effect. We chose seven methods to evaluate: propensity score matching (PSM), inverse probability of treatment weighting (IPTW) (Rosenbaum & Rubin, 1983), outcome regression (OR), BART (Chipman et al., 2007), causal forests (CF) (Wager & Athey, 2017), doubly-robust estimation (DRE) (Funk et al., 2011), and a neural network-based method (NN) (Shi et al., 2019). As a baseline for comparison, we also included a naive method that simply estimates $E[Y|T = 1] - E[Y|T = 0]$ from the observational data. Details about these methods can be found in the supplementary material.

5.3. Experiments

For the RCT, APO and simulator data sets[5], we selected a single treatment and outcome pair, and selected a biasing covariate that was pre-treatment and correlated with outcome. For each data set, we calculated the unbiased ATE to use as ground truth. We then sub-sampled, as shown in Algorithms 1 and 2, to produced a biased data set. All algorithms were applied to the biased data, producing estimates of ATE.This process was repeated for 30 trials. For data sets with more than 2,000 individuals, we sub-sampled to 2,000, to keep sample sizes comparable between data sets.

For data sets with a binary outcome, we used risk difference instead of ATE, calculated as $P(Y = 1|do(T = 1)) - P(Y = 1|do(T = 0))$. Error in ATE and risk difference estimation was calculated as the absolute difference between the predicted value and the ground truth value calculated on the unbiased data. ATE and outcome estimates were normalized by the range of the outcome variable for easier comparison.

It is important to note that, while the biasing procedure was the same for all the RCT, Simulator, and APO data sets, the strength of the bias may differ substantially. This is because two factors influence the strength of the confounding bias: (1) the strength of the treatment sampling function $f(C^b)$; and (2) the strength of the latent dependence between the biasing covariates and outcome. In order to ensure that confounding bias was introduced, we only chose biasing covariates that were correlated with outcome. However, the strength and nature of this dependence can vary significantly.

[5]The synthetic-response data sets were, by design, pre-biased, so no additional processing was necessary

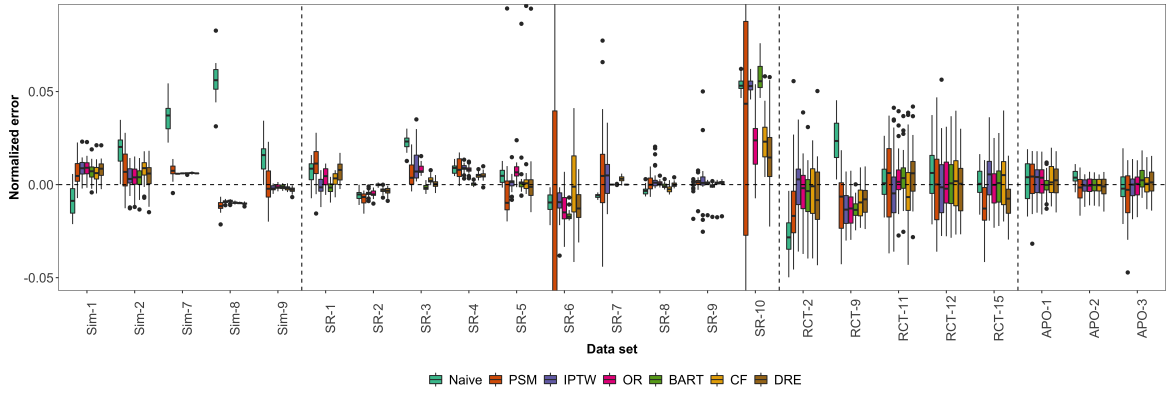

Figure 3. Normalized error in estimating ATE for data sets with continuous outcome. Sim denotes simulator, SR denotes synthetic-response data sets, RCT denotes randomized controlled trials and APO denotes computational systems.

This means that even if we induce strong dependence in the sampling procedure, the overall confounding bias may be weak, due to a weak effect of the biasing covariates on outcome. To assess how much confounding bias is introduced in each data set, the naive algorithm (described above) serves as a simple estimate of that bias, by showing the error induced by not accounting for the biasing covariates.

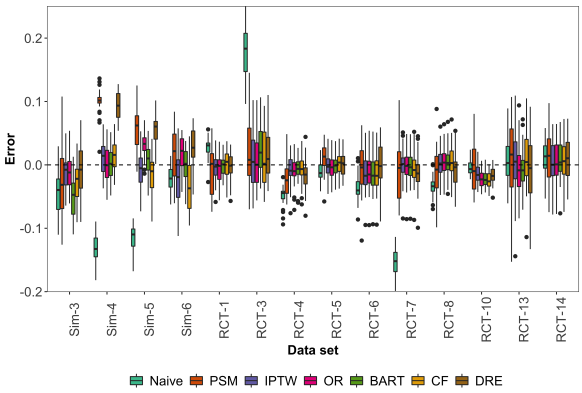

Figure 4. Error in estimating risk difference for data sets with binary outcomes.

5.4. Results

Effect estimation results for the experimental setup described above are shown in Figure 3 for data sets with continuous outcomes and Figure 4 for data sets with binary outcomes. These results can be analyzed in two main ways: comparing the performance of different algorithms within each data set, and comparing how algorithm performance differs between different data sets. Most evaluations in the literature focus solely on comparing performance within individual data sets. Evaluating across data sources, though, can also provide a useful understanding of how different data design choices affect estimates of relative performance.

All methods perform well overall, with error typically centered around zero. Propensity score matching consistently has the highest variability. This is consistent with the literature, which shows that the pruning done by propensity score matching can increase data set imbalance, and thus increase estimation bias, when matching solely on the propensity score (King & Nielsen, 2019).

While performance between the methods is very similar for most for the RCT data sets, the synthetic-response data sets show substantially more variability between different algorithms. For example, IPTW has higher variability on some synthetic-response data sets, BART performs well overall, and BART and IPTW perform poorly on SR-10.

One interesting feature of these results is that, overall, performance between the RCT and APO data is fairly similar, with similar variability ranges and most methods performing about the same. The simulators have lower variability in general, but, for the most part, have similarly equivalent performance across methods. This stands in contrast to the synthetic-response data sets, where we see far more variability between methods on the same data set.

The similarity between the RCT and APO data is a good sign. APO data can be thought of as an ideal, but hard to come by, situation. In practice, we are far more likely to only observe one potential outcome for every individual, as is the case with RCTs. Thus, OSRCT allows us to get nearly identical results using data that is far easier to acquire than OSAPO.

This contrast between the synthetic-response data sets and the other three types has several possible explanations. One is that the complexity of the response surface in the synthetic-response data is far higher than that of the other data sets. Given that the response surface in the RCT data sets arise naturally in real-world systems, this suggests that the level of complexity in the synthetic-response data sets is

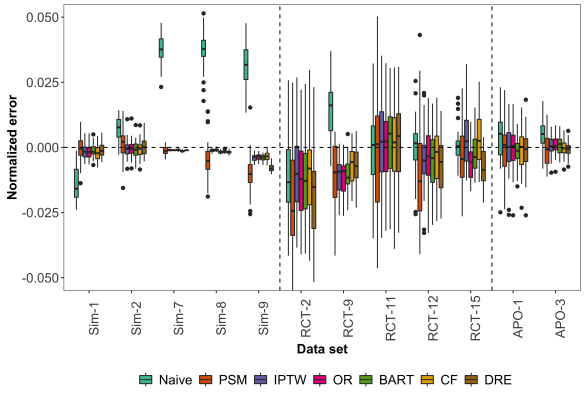

Figure 5. Normalized error in estimating ATE with two biasing covariates, for data sets with continuous outcome

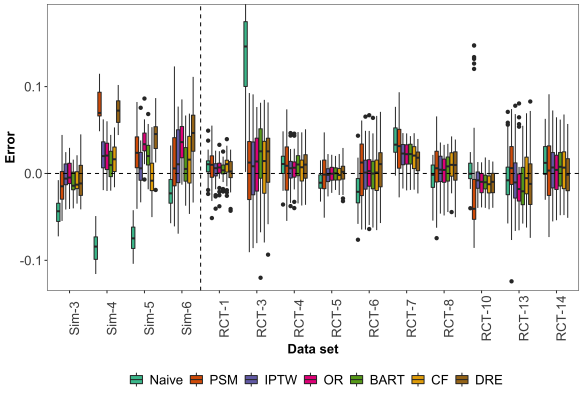

Figure 6. Error in estimating risk difference with two biasing covariates, for data sets with binary outcome

not realistic.

Another possible explanation is that the treatment assignment in the RCT data is very simplistic (i.e., based on the value of only a single biasing covariate), while the treatment in the synthetic-response data sets is assigned based on a complex combination of many covariates. To test this hypothesis, we defined a more complicated biasing function, using a combination of two covariates that are correlated with outcome. Where possible, numeric covariates were chosen. However, some data sets have only factor covariates, or a very limited number of numeric covariates, so a mix of factor and numeric biasing covariates was used.

Results with two biasing covariates are shown in Figures 5 and 6. For the most part, estimates are similar to those produced with a single biasing covariate, and we still do not see the differences between algorithms that we do for the synthetic-response data sets. It is possible that an even more complicated biasing function, potentially with many more covariates, is necessary for these results to be similar. However, if the complexity of the treatment function is the cause of the performance difference, we would expect methods that focus on treatment modeling (such as IPTW and propensity score matching) to be performing consistently worse in this regime, which is not what we observe.

In summary, we demonstrate that three types of evaluation data (simulators, computational APO data sets, and RCTs) all produce results that are markedly different from the results produced by evaluating on synthetic-response data. Specifically, the results from these three types of evaluation data suggest that many modern methods for estimating causal effects are roughly equivalent in performance on realistic causal inference tasks.

6. Conclusion

Research progress in machine learning has long depended on high-quality empirical evaluation. Research in causal inference has been hindered due to sparse empirical data resources. The growth in such data resources is slow, and the breadth of such data is still limited, especially when compared to the wealth of evaluation data sets available for associational machine learning.

Data from RCTs provides a large and growing source of data that can be used to evaluate causal inference methods. RCT data has been widely collected by researchers in many fields over many years and is increasingly being made available for wider use. Harnessing this wealth of data with OSRCT can substantially increase what we know about the absolute and relative performance of causal inference methods.

Acknowledgments

Thanks to Justin Clarke for help preparing data sets and to Kaleigh Clary, Sam Witty, and Kenta Takatsu for their helpful comments. This research was sponsored by the Defense Advanced Research Projects Agency (DARPA), the Army Research Office (ARO), and the United States Air Force under under Cooperative Agreement W911NF-20-2-0005 and contracts FA8750-17-C-0l20 and HR001120C0031. Any opinions, findings and conclusions or recommendations expressed in this document are those of the authors and do not necessarily reflect the views of DARPA, ARO, the United States Air Force, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation herein.

References

Arceneaux, K., Gerber, A. S., and Green, D. P. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, pp. 37–62, 2006.

Atan, O., Jordon, J., and van der Schaar, M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 2021.

Banzi, R., Canham, S., Kuchinke, W., Krleza-Jeric, K., Demotes-Mainard, J., and Ohmann, C. Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials*, 20, 2019.

Chipman, H. A., George, E. I., and McCulloch, R. E. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems*, pp. 265–272, 2007.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34:43–68, 2019.

Drazen, J. M. Sharing individual patient data from clinical trials. *New England Journal of Medicine*, 372(3):201–202, 2015.

Dryad, 2020. URL https://datadryad.org/stash/. [Online; accessed 3-June-2020].

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011.

Gentzel, A., Garant, D., and Jensen, D. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems 32*, pp. 11722–11732. 2019.

Godlee, F. and Groves, T. The new BMJ policy on sharing data from drug and device trials. *British Medical Journal*, 345, 2012.

Guillaume, F. and Rougemont, J. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 22:2256–2557, 2006.

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

Hill, J. L., Reiter, J. P., and Zanutoo, E. L. A comparison of experimental and observational data analyses. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, 2004.

Jaciw, A. P. Assessing the accuracy of generalized inferences from comparison group studies using a within-study comparison approach: The methodology. *Evaluation Review*, 40(3):199–240, 2016.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pp. 9269–9279, 2018.

Kallus, N., Mao, X., and Udell, M. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems 31*, pp. 6921–6932. 2018a.

Kallus, N., Puli, A. M., and Shalit, U. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pp. 10888–10897, 2018b.

King, G. and Nielsen, R. A. Why propensity scores should not be used for matching. *Political Analysis*, 27(4), May 2019.

Kuntz, R. E., Antman, E. M., Califf, R. M., Ingelfinger, J. R., Krumholz, H. M., Ommaya, A., Peterson, E. D., Ross, J. S., Waldstreicher, J., Wang, S. V., et al. Individual patient-level data sharing for continuous learning: A strategy for trial data sharing. *NAM Perspectives*, 2019.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pp. 604–620, 1986.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth*

ACM International Conference on Web Search and Data Mining, pp. 297–306, 2011.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Mary, J., Preux, P., and Nicol, O. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *International Conference on Machine Learning*, pp. 172–180, 2014.

Miller, J., Hsu, C., Troutman, J., Perdomo, J., Zrnic, T., Liu, L., Sun, Y., Schmidt, L., and Hardt, M. Whynot, 2020. URL https://doi.org/10.5281/zenodo.3875775.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Morgan, S. L. and Winship, C. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

NIH National Institute on Drug Abuse Data Share Website, 2020. URL https://datashare.nida.nih.gov/. [Online; accessed 3-June-2020].

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., Karam, G., Kalra, D., Knowles, R. L., Krleža-Jerić, K., Kubiak, C., Kuchinke, W., Kush, R., Lukkarinen, A., Marques, P. S., Newbigging, A., O'Callaghan, J., Ravaud, P., Schlünder, I., Shanahan, D., Sitter, H., Spalding, D., Tudur-Smith, C., van Reusel, P., van Veen, E.-B., Visser, G. R., Wilson, J., and Demotes-Mainard, J. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open*, 7(12), 2017.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62 (3):54–60, 2019.

Request for Public Comment on DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance, 2019.

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

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, April 2005.

Saito, Y. and Yasui, S. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pp. 8398–8407. PMLR, 2020.

Salganik, M. J., Dodds, P. S., and Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

Shadish, W. R., Clark, M. H., and Steiner, P. M. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344, 2008.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Shi, C., Blei, D. M., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2503–2513, 2019.

Shimoni, Y., Yanover, C., Karavani, E., and Goldschmnidt, Y. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Suvarna, V. R. Sharing individual patient data from clinical trials. *Perspectives in Clinical Research*, 6:71–72, 2015.

Tang, L., Rosales, R., Singh, A., and Agarwal, D. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1587–1594, 2013.

Tang, L., Jiang, Y., Li, L., and Li, T. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 73–80, 2014.

The Knowledge Network for Biocomplexity, 2020. URL https://knb.ecoinformatics.org/. [Online; accessed 3-June-2020].

Tu, R., Zhang, K., Bertilson, B., Kjellstrom, H., and Zhang, C. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In *Advances in Neural Information Processing Systems 32*, pp. 12793–12804. 2019.

UK Data Service, 2020. URL https://ukdataservice.ac.uk/. [Online; accessed 3-June-2020].

University of Michigan Institute for Social Research, 2020. URL https://www.icpsr.umich.edu/web/pages/. [Online; accessed 3-June-2020].

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017.

Witty, S., Takatsu, K., Jensen, D., and Mansinghka, V. Causal inference using Gaussian processes with structured latent confounders. In *International Conference on Machine Learning*, pp. 10313–10323. PMLR, 2020.

Yale Institution for Social and Policy Studies Data Archive, 2020. URL https://isps.yale.edu/research/data. [Online; accessed 3-June-2020].

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

Zeng, C., Wang, Q., Mokhtari, S., and Li, T. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2025–2034, 2016.

Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12207–12215, 2021.