June 2021

Improving Evaluation Methods for Causal Modeling

Amanda Gentzel
University of Massachusetts Amherst

IMPROVING EVALUATION METHODS FOR CAUSAL MODELING

A Dissertation Presented

by

AMANDA GENTZEL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2021

College of Information and Computer Sciences

IMPROVING EVALUATION METHODS FOR CAUSAL MODELING

A Dissertation Presented

by

AMANDA GENTZEL

Approved as to style and content by:

David Jensen, Chair

Ben Marlin, Member

Justin Gross, Member

Madalina Fiterau, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

ACKNOWLEDGMENTS

I would not be where I am without the my advisor, David Jensen. I feel like I stumbled into this field, with no knowledge or experience, and I am forever grateful that he welcomed me into his lab. Through my time in KDL, David has taught me so much about how to approach new problems and how to share and discuss ideas, all with a positive and curious energy that was infectious. His patience and support have been helpful beyond words, especially during those first few years when I barely spoke at meetings and feared I wasn't cut out for this. I've learned and improved so much thanks to his guidance, and I am forever grateful for everything he has done for me. I also need to thank my undergraduate advisor, David Shaffer, and undergraduate math teacher, Carolyn Cuff, for pushing me to do REUs and consider graduate school; without them, I would not have even known that this was an option.

I am also grateful for the amazing group of labmates I've had the pleasure of working with over the years. Coming from a very small computer science program in undergrad, it was amazing, and frequently intimidating, to be surrounded by such a thoughtful and friendly group of researchers. I have learned so much from all of you: Pracheta Amaranath, David Arbour, Akanksha Atrey, James Atwood, Kate Avery, Lissa Baseman, Javier Burroni, Erica Cai, Justin Clarke, Kaleigh Clary, Dan Corkill, Lisa Friedland, Reilly Grant, Jack Kenney, Phil Kirlin, Marc Maier, Katerina Marazapoulou, Hüseyin Oktay, Purva Pruthi, Matt Rattigan, Kenta Takatsu, Brian Taylor, Emma Tosch, Sankaran Vaidyanathan, Sam Witty, and Andy Zane. I also need to thank Deb Bergeron, for making all of the administrivia of grad school that much smoother, and Leeanne Leclerc, whose deep knowledge and wizardry with the CICS and grad school requirements was endlessly appreciated.

I also have to thank all of the friends I've had here, who have formed an indispensable support network. The friendliness and strong community within the CS program, especially during the first half of my PhD, was amazing. Without their support and friendship, I doubt I would have finished this. Lissa Baseman, who joined me in paper airplane-throwing in the lab and Juston Moore, who helped carry me through Algorithms; Kevin Winner and Kaleigh Clary, with whom I've played many games and whose friendship means more to me than they can know; Keen Sung, Larkin Flodin, and Janet Guo, who helped maintain my sanity for this final pandemic year of grad school; all of my roommates over the years, especially Chelsea Marcho, Laura Titrud, Lissa Baseman, Kevin Winner, Keen Sung, Larkin Flodin, Myungha Jang, Tony Ohmann, Li Yang Ku, Joie Wu, and Alan Sempruch: John Vilk, Dirk Ruiken, Zack Serritella, Erin McVicar, and Chris Ibbotson, who joined me in epic D&D adventures; and the other wonderful friends I've made through UMass: Luis Pineda, Shiri Dori-Hacohen, Addison Mayberry, Garrett Bernstein, Albert Williams, Roy Adams, Tiffany Liu, Elizabeth Do, J.D. DeVaughn-Brown, Kevin Spiteri, Presley Pizzo, Stephen Oloo, Keiko Konoeda, and Dan O'Shea. I also need to thank my non-UMass friends: Laura Titrud, one of the first friends I made in the area and with whom I've shared many tacos and episodes of Glee; and the X9 Games crew, who I will sorely miss crushing in board games, with special mention to Matt Duval, Tim Bennett, Bryan Denley, Katie Williams, Gwen Provost, Greg Lyon, and Alicia Brody. All of them have been an indispensable support network

Finally, I need to thank my family, for their unwavering support for my hairbrained decision to move to Massachusetts and get a PhD. My parents Margaret and David, for always being there for advice and support, and never (well, rarely) asking me when I was going to finish; my sisters Jen and Becca, for the countless nights spent gaming, and countless days spent talking; my grandparents, all of whom I sorely miss, but know they would be very proud of me; and my aunts and uncles and cousins,

who are such a positive force that I always look forward to seeing. It takes a village, and I don't think I could have asked for a better one. Thank you.

ABSTRACT

IMPROVING EVALUATION METHODS FOR CAUSAL MODELING

MAY 2021

AMANDA GENTZEL

B.Sc., B.M., WESTMINSTER COLLEGE

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor David Jensen

Causal modeling is central to many areas of artificial intelligence, including complex reasoning, planning, knowledge-base construction, robotics, explanation, and fairness. Active communities of researchers in machine learning, statistics, social science, and other fields develop and enhance algorithms that learn causal models from data, and this work has produced a series of impressive technical advances. However, *evaluation techniques* for causal modeling algorithms have remained somewhat primitive, limiting what we can learn from the experimental studies of algorithm performance, constraining the types of algorithms and model representations that researchers consider, and creating a gap between theory and practice. We argue for expanding the standard techniques for evaluating algorithms that construct causal models. Specifically, we argue for the addition of evaluation techniques that use *interventional measures* rather than structural or observational measures, and that

evaluate with those measures on *empirical data* rather than synthetic data. We survey the current practice in evaluation and show that, while the evaluation techniques we advocate are rarely used in practice, they are feasible and produce substantially different results than using structural measures and synthetic data. We also provide a protocol for generating observational-style data sets from experimental data, allowing the creation of a large number of data sets suitable for evaluation of causal modeling algorithms. We then perform a large-scale evaluation of seven causal modeling methods over 37 data sets, drawn from randomized controlled trials, 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. This difference demonstrates the importance of using data from a variety of sources when evaluating any causal modeling methods.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

INTRODUCTION

Causal modeling is a rapidly growing field within computer science, relevant to many areas including complex reasoning, planning, and robotics. It also has a long history of use in other fields, including economics, social science, and epidemiology. Causal reasoning is at the core of all experimental science. While causal modeling has only recently become popular among computer science researchers, these other fields have long recognized the value that causal modeling provides.

Causal models have several advantages over strictly associational models.

- **Prescriptive:** Knowledge of the effects of interventions can guide decision making. For example, while it can be interesting to know that a public policy is correlated with a change in economic outcome for a population, unless we know that the policy is the cause of the change in outcome, we cannot recommend the policy with any confidence.

- **Predictive:** A causal model allows us to query what the outcome will be if a certain action is taken. By intervening in a causal model, we can estimate what the outcome will be if we make the corresponding intervention in the real world.

- **Robust:** A causal model is able to represent the mechanistic dependencies between variables, rather than just the association. If the distribution of some variables changes, but the mechanisms remain the same, then a causal model should be robust to this change and still provide accurate estimates.

- **Explanatory:** With the increasing desire to understand the decisions made by machine learning systems, causal modeling is uniquely qualified for this task.

A causal model facilitates counterfactual queries, questions of the form "What would have happened if...", allowing for reasoning about the underlying causal mechanisms of a causal system.

Because of these advantages, interest in causal modeling within computer science has grown significantly in recent years. Due to the large amount of observational data available, and the difficulty and cost of performing experiments, many methods have been created to learn causal models from observational data. These approaches use a variety of techniques to disentangle causality from mere association and have had some success. A clear understanding of how well these methods work, however, has been hampered by the difficult question of how to evaluate them.

While ground truth is easily available for testing statements of association, testing statements of causality is significantly harder. While there are plenty of real-world data sets available for evaluating predictive performance, data for evaluating causal models has much stricter requirements. Without data from intervening in a system, or knowledge of ground truth causal dynamics, we cannot assess whether an algorithm has learned the correct causal dependencies. In addition, most causal modeling algorithms are designed to run on observational data, acquired from observing a system rather than acting in it. Without data that is observational, but where the results of interventions are known (to provide ground truth causal dependencies), traditional causal modeling algorithms can not be effectively evaluated, and most data that is available right now does not meet this requirement.

To assess the correctness of a causal model, the model should be evaluated on how well it captures actual causal effects in real-world data. However, the difficulty in acquiring such data for evaluation has lead to the widespread adoption of alternative evaluation techniques that fall short of this. Synthetic data is used frequently, to get around the need for ground truth causal knowledge. When real-world data is used, researchers often use it as merely a proof of concept with no ground truth; in this

situation, the only evaluation available consists of looking at the learned dependencies and making sure they "seem reasonable."

Sometimes, even when ground truth causal direction is known (because the observational data is from a system with known dynamics), the magnitude of that dependence is not. This focus on structure alone has lead to the widespread adoption of measures that only assess whether the correct structure has been learned, largely ignoring whether the strength of the dependencies is correct as well. Even in those cases where researchers evaluate their algorithm on real-world data with known causal dependencies, they are generally only able to evaluate with one or two such data sets, providing a limited assessment of the general effectiveness of proposed algorithms.

To address these problems and gaps, this work makes several contributions. Specifically, we:

1. Identify the interconnected nature of the components of an evaluation and provide a useful decomposition;

2. Perform an extensive survey of current practice in evaluation for causal modeling;

3. Empirically demonstrate the limitations of current evaluation methods and the value of evaluating causal modeling algorithms using empirical data and interventional measures;

4. Specify an algorithm for converting an ideal interventional data set into a pseudo-observational data set, suitable for evaluating causal modeling algorithms;

5. Prove that a similar algorithm can be applied to data from randomized controlled trials, and show how data discarded during sampling can be used to evaluate outcome estimation; and

6. Perform a large-scale evaluation of current causal modeling algorithms with greater diversity of data set types than any previous evaluation.

The remainder of the document proceeds as follows:

- **Chapter 1** discusses background on causal modeling, describing methods that focus on multivariate structure learning and bivariate effect estimation.
- **Chapter 2** discusses related work, outlining the components of an evaluation and popular current evaluation methods.
- **Chapter 3** discusses work on evaluating causal modeling methods, surveying current practice and making recommendations for how evaluation should be performed.
- **Chapter 4** describes how data from randomized controlled trials can be used for evaluation.
- **Chapter 5** reports the results of a large-scale evaluation of seven causal modeling methods, using 37 data sets from a variety of sources.

CHAPTER 1

BACKGROUND

While causality has been defined in many different ways, the definition we use is focused on the concept of *intervention* or *manipulation*. Cook and Campbell [45] state that: "The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. ... Causation implies that by varying one factor, I can make another vary."

The focus of this work is on algorithms for causal modeling. Many algorithms have been developed to estimate causal model structure and the strength of causal dependence. In this chapter, we describe common representations and approaches for causal modeling. These techniques are actively used in practice, and we will later assess how the results of these techniques can be evaluated. Note that in the literature, the terms 'causal learning', 'causal modeling', 'causal inference', and 'causal discovery' are frequently used interchangeably or with subtle and inconsistently applied shades of meaning. Here, we use the term 'causal modeling' to refer to the group of models and methods used to infer the structure and strength of causal dependence.

1.1 Causal Terminology

Causal dependencies are often discussed in terms of *treatments* and *outcomes*. This terminology is borrowed from the medical community, with *outcome* representing the quantity of interest (e.g., some measure of patient health) and *treatment* representing an action that can be taken to possibly affect outcome (e.g., taking a drug). Other possibly relevant variables (e.g., age, health history) are called *covari-*

ates or *confounders*. Depending on the situation, there may be many treatments, outcomes, and covariates. A standard distinction, though, is that treatments are the variables that can be manipulated, outcomes are the post-treatment variables of interest, and covariates are additional measured variables that may affect treatment and/or outcome.

As mentioned above, we define causality in terms of manipulation. A treatment X is said to cause an outcome Y if manipulating the value of X results in a change in the probability distribution of Y. This notion is formalized with the concept of an intervention as described by Pearl's do-calculus [150]. The do-calculus is a framework that allows for reasoning about the effects of interventions in graphical models. The operator *do* represents a manipulation of the network. Performing $do(X = x)$ sets the value of the variable X to a specific value x. This operation modifies the graph structure, removing all incoming edges to node X and forcing X to take the value x.

1.2 Causal Graphical Models (CGMs)

Causal graphical models are a class of models used to encode causal information about a set of variables. A few types of graphical models are commonly used for causal modeling, including directed acyclic graphs [72] and chain graphs [148]. Our focus will here will be on directed acyclic graphs, since they are the most prevalent in the computer science community.

Directed acyclic graphs (or DAGs) are a class of graph that (as the name suggests) is both directed and acyclic. For nodes A and B, if an edge $A \to B$ exists, then A is called a *parent* of B, and B is called a *child* of A. There is a directed path from A to B if there exists a sequence of nodes, starting at A and ending at B, where the child of each node is the same as the parent of the next one. A cycle occurs when there exists a directed path from any node to itself. Acyclicity requires that there are no directed cycles in the graph.

A *directed graphical model* is a DAG with probabilistic semantics [29]. A network with random variables $A_1, A_2, A_3, ..., A_n$ represents the joint probability distribution $P(A_1, A_2, A_3, ..., A_n)$. However, this probability distribution can be factored, as represented by the DAG structure, into a form that is far simpler to compute. Directed graphical models satisfy the *local Markov property*, which states that each variable is conditionally independent of its non-descendants given its parents. This allows the joint distribution to factor into the probabilities of each variable conditioned on its parents:

$$P(A_1, ..., A_n) = \prod_{A_i \in A} P(A_i | Pa(A_i)) \tag{1.1}$$

Directed graphical models are often used to represent causal dependence, though they inherently do not require that edges represent causality. When a directed graphical model is given a causal interpretation, we call it a *causal graphical model (CGM).* In a causal graphical model, directed edges are interpreted as causal dependence, rather than just probabilistic dependence. Hernan and Robins [86] offer this definition of a causal DAG:

Definition 1.2.1. A causal DAG is a DAG in which (1) the lack of an arrow from node V_j to V_m can be interpreted as the absence of a direct causal effect of V_j on V_m (relative to the other variables on the graph), (2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph, and (3) any variable is a cause of its descendants.

1.3 Approaches to Causal Modeling

Many research communities, including computer science, economics, and statistics, have developed techniques for inferring causal dependence from observational data. Since the focus of these communities is different, the goal of causal modeling techniques varies among them. We will divide these approaches primarily based on

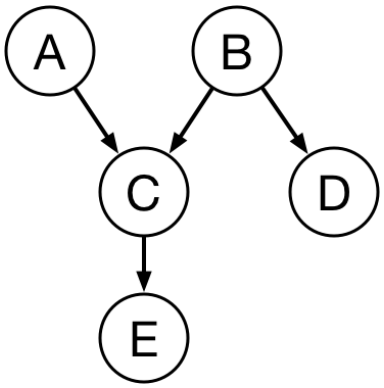

Figure 1.1. Example directed acyclic graph

two aspects: whether they aimed at learning structure or parameters, and whether they operate on bivariate or multivariate data.

1.3.1 Multivariate Structure Learning

Much of the work in computer science on causal modeling focuses on learning causal graphical models. Several classes of algorithms have been created to learn the structure of these models from observational data. These algorithms are designed for data sets consisting of independent and identically distributed (i.i.d.) samples, and the goal is to learn the structure of causal dependence among the variables.

Constraint-based algorithms use conditional independence tests to infer which edges exist in the model underlying the data. First, they use these tests to learn a skeleton of undirected edges between the variables. They then apply a series of orientation rules to determine the direction of edges. Using the results of conditional independence tests alone, it is not always possible to fully orient all edges, so a set of possible structures, referred to as a *Markov equivalence class*, is returned instead. After the structure is learned, parameters can be fit to any member of the Markov equivalence class. Popular constraint-based algorithms include PC [179], Grow-Shrink [134], Incremental Association (IAMB) [190], and FCI [179].

Most constraint-based learning algorithms make two standard assumptions: causal sufficiency and faithfulness. *Causal sufficiency* states that all possible common causes of measured variables are included in the analysis [179]. This assumption prevents, for example, the existence of a latent variable that causes both A and B in Figure 1.1, but would allow for latent variables that cause either A or B.

Faithfulness states that all independencies that are present in the data match those of the underlying causal graph [179]. For example, in Figure 1.1, C and D are dependent because they share B as a parent. However, it is possible that, in data generated from this structure, C and D happen, by chance, to be independent. This would be a violation of faithfulness, because the distribution of the generated data is not faithful to the model. In this example, current techniques are unable to infer, from the data alone, that C and D are not independent, making it impossible to learn the true causal structure.

The *PC algorithm* [179] is one of the most popular constraint-based methods. It starts with a fully-connected graph and iteratively performs conditional independence tests on pairs of variables, with an increasingly large separating set. If a pair of variables is ever found to be conditionally independent, the edge between them is removed. This results in an undirected skeleton. The algorithm then orients as many edges as it can, using rules implied by d-separation (a set of graphical criteria for determining independence facts from a graphical model) and the acyclicity constraint. This results in a Markov equivalence class of DAGs, corresponding to all legal orientations of the undirected edges. The FCI algorithm is similar to the PC algorithm, but it does not assume causal sufficiency.

Score-based algorithms use heuristics to find the structure that optimizes some score function. Popular score-based algorithms include Greedy Equivalence Search (GES) [41] and greedy hill-climbing search [62].

Hybrid algorithms combine aspects of both constraint-based and score-based methods. For example, the max-min hillclimbing (MMHC) algorithm uses an initial constraint-based pass to learn a skeleton and then uses a score-based approach to orient the edges [191]. Other variants of structure learning algorithms may incorporate different types of data or background knowledge, such as providing initial temporal ordering information or allowing for a limited number of real-world experiments [50, 47, 205].

1.3.2 Bivariate Orientation

In some domains, rather than aiming to determine the causal structure of a large set of variables, the focus is to determine the structure between two variables (i.e., the existence and direction of a single edge). In many cases, the existence of dependence between the two variables is assumed or already determined, so the goal is only to determine the orientation of that edge.

In economics, **Granger causality** [77] is commonly applied to this problem. Granger causality operates on time series data, making use of the intuition that a cause should be useful in forecasting future values of its effect. A simple test for whether X Granger causes Y involves running two regressions,

$$y_t = a_0 + \sum_{i=1}^{t-1} a_i y_i + e_t \tag{1.2}$$

and

$$y_t = a_0 + \sum_{i=1}^{t-1} a_i y_i + \sum_{j=1}^{t-1} b_j x_j + e_t \tag{1.3}$$

If the residuals of these two regressions are significantly different, then lagged values of X are providing information about Y_t that is not contained in lagged values of Y_t, so X is said to Granger cause Y. These tests are typically performed in both directions (constructing pairs of regression models for both the effect of X on Y and

the effect of Y on X), and whichever direction produces the strongest effect is said to be the direction of Granger causality. For the conclusions drawn from a Granger causality test to be correct, additional covariates should be included in the regression to control for any possible confounders.

Additive noise models make assumptions about the distribution of the dependence between the two variables (generally, that they follow a structural equation model with additive noise) and exploit the asymmetry in the noise to determine the likely direction of causal dependence [90]. In the basic formulation, the dependence between x and y is represented by the generative model $y := f(x) + n$, where x and n are Gaussian and statistically independent. Hoyer et al. show that, in most cases, the shape of the conditional densities $P(y|x)$ and $P(x|y)$ are different. When this is the case, assuming the variables are non-independent, we can test the fit of the regressions,

$$y := f(x) + n \tag{1.4}$$

and

$$x := g(y) + n \tag{1.5}$$

If, for example, the fit of $y := f(x) + n$ is rejected and $x := g(y) + n$ is not, we have evidence that the true direction is $Y \rightarrow X$.

1.3.3 Bivariate Effect Estimation

Bivariate effect estimation methods are used when the direction of causality is known in advance (for example, a drug trial, where the drug is assumed to have some effect on patient outcome). The goal is to estimate the level of causal effect from treatment to outcome, possibly controlling for a set of known confounders.

The **potential outcomes framework** [163] provide a framework for reasoning about dependence that makes use of the concept of the counterfactual. Suppose there is an observational study that aims to determine the effect of a binary treatment, T,

on outcome, Y. To determine the causal effect across the whole population, many methods calculate the average treatment effect (ATE):

$$ATE = E[Y^1] - E[Y^0] \tag{1.6}$$

where Y^1 and Y^0 are the individual outcomes for $T = 1$ and $T = 0$, respectively. In practice, however, we can rarely observe outcomes for both $T = 1$ and $T = 0$ for the whole population. For example, if an individual i was assigned to treatment 1, then we never observe Y_i^0 — referred to as the counterfactual outcome, the outcome we would observe if, contrary to fact, i were assigned treatment 0. When a randomized experiment is performed, the populations that receive each value of treatment are assumed to be equivalent, so ATE can be calculated using only the groups that receive each value of treatment. However, in many circumstances, randomized experiments are not possible. Potential outcomes provides a framework for reasoning about both potential outcomes of treatment. Techniques that are used by the potential outcomes community to estimate causal effects include exact matching methods, propensity score matching methods, and inverse probability of treatment weighting.

Propensity-score matching aims to form matched pairs of individuals in the treatment and control groups that have the same probability of receiving treatment. By matching pairs that were equally likely to have been treated (but in which one received treatment and one did not), we can estimate the differences due to the treatment alone rather than due to confounders. Propensity score methods estimate a propensity score, $P(T = 1|X_1, ..., X_n)$, for covariates $X_1, ..., X_n$. Individuals in the treated population can then be matched to individuals in the control population with similar conditional probability of treatment, and ATE can be calculated as the difference in outcome between these two sub-populations. Because the typical implementation of propensity-score matching matches the control population to the treated population, the resulting estimate is actually the average treatment effect

on the treated (ATT), which is equivalent to ATE when the treated and control populations are the same [158].

Inverse probability of treatment weighting (IPTW) is similar to propensity score matching, in that both estimate the probability of treatment and use that to control for confounding. Rather than using the probability of treatment to match individuals between the treatment and control populations, IPTW *weights* the outcomes of every individual according to their probability of treatment and uses these weighted outcomes to estimate ATE [158]. For treatment $T \in \{0, 1\}$, IPTW calculates ATE as

$$ATE = E[\frac{YT}{p(T|X)}] - E[\frac{Y(1-T)}{1-p(T|X)}] \tag{1.7}$$

Outcome regression is one simple approach for effect estimation that models outcome given treatment and all measured covariates. Unlike the potential outcomes approaches discussed above, outcome regression makes no attempt to model the treatment mechanism, focusing solely on effectively modeling outcome. Recent studies have suggested that effectively modeling outcome may be more important than trying to account for differences in treatment assignment [53].

Bayesian Additive Regression Trees (BART) use a tree-based model to estimate the response surface, allowing for estimation of both ATE and individual outcomes [43]. Regression trees are a type of tree used when the outcome is continuous, which partition the input data into subgroups with similar outcomes. BART creates an ensemble of sequentially-learned regression trees, with a regularization prior to keep the effects of individual trees small. Estimates for the ensemble are obtained by summing the outputs of all the trees. When used for causal modeling, all observed covariates and treatment are used as predictors of outcome, and estimates of ATE can be obtained by estimating outcome for all individuals with both $T = 1$ and $T = 0$ and calculating the mean difference. This methods, similar to outcome regression, focuses solely on modeling outcome. However, it is common to include an

estimate of the propensity score as an additional covariate to account for treatment effect heterogeneity [88].

The above methods focus on modeling either treatment or outcome. However, if such models are misspecified, the effect estimate can be biased. **Doubly-robust methods** are designed to avoid this issue, producing an unbiased estimate of ATE as long as *either* the treatment or the outcome model is correctly specified. This is commonly implemented as a combination of IPTW weighting and outcome regression [60]. For treatment T, covariates X, outcome Y, propensity score $\pi(X)$, and outcome estimate $\hat{Y}$, we can define a doubly robust estimate of ATE as

$$ATE = \frac{1}{n}\sum_{i=1}^{n}[\frac{T_i Y_i}{\pi(X_i)} - \frac{(T_i - \pi(X_i))\hat{Y}_i^1}{\pi(X_i)}] - \frac{1}{n}\sum_{i=1}^{n}[\frac{(1-T_i)Y_i}{\pi(X_i)} - \frac{(T_i - \pi(X_i))\hat{Y}_i^0}{\pi(X_i)}] \tag{1.8}$$

where $\hat{Y}_i^1$ and $\hat{Y}_i^0$ represent the estimates of $E[Y|T=1, X_i]$ and $E[Y|T=0, X_i]$, respectively.

The second half of each term of the difference can be thought of as an adjustment to the IPTW-adjusted first term. If either $\pi(X_i)$ or $\hat{Y}_i$ is correctly specified, in the sample limit, the second term zeroes out, leading to an unbiased estimate of $E[Y^1] - E[Y^0]$ [60].

Unlike traditional random forests which estimate the value of an outcome variable, **causal forests** are random forests that specifically estimate ATE [193]. They make use of causal trees [23], which estimate ATE at the leaf nodes by splitting such that the the number of training points at the leaf node is small enough to be treated as though they came from a randomized experiment. A causal forest then averages the ATE estimates from the causal trees in the ensemble to get an overall estimate of ATE. This approach has similarities to matching methods, where individuals are partitioned into matched groups at leaf nodes.

While **causal graphical models** are most frequently used to learn structure, they can also be used for effect estimation. Pearl's do-calculus defines manipulations of the graph that can allow for estimation of the effect an intervention. Pearl also defines a set of adjustment methods, the back-door and front-door formulas, that can be used to estimate causal effects in graphical models [151].

Shi et al. [173] propose a neural-network-based method, using a new proposed architecture called **Dragonnet**. This approach uses a deep neural network to produce a representation layer of the covariates. This representation layer is then used to predict both treatment and outcome. The prediction of treatment acts as a propensity score, which is used to adjust for confounding when estimating treatment effect. Dragonnet net is one example of a class of neural-network-based approaches for causal modeling, which generally follow a similar approach [99, 170, 167, 128, 206].

CHAPTER 2

RELATED WORK

As already mentioned, many different research communities have developed methods for constructing causal models from observational data. However, assessing the absolute and relative performance of these methods requires some evaluation method. For ease of discussion, we decompose evaluation methods into three key components:

- **Data source** — The data provided to the causal modeling algorithm
- **Algorithm** — The causal modeling algorithm under evaluation
- **Evaluation measure** — A descriptive or numerical measure that is used to assess the causal inferences of the model produced by the algorithm

These dimensions are highly dependent; a choice of one can determine feasible choices for the other two. For example, models learned from observational macroeconomic data often cannot be compared against a known structure because there exists no ground truth, and models consisting only of non-parameterized structure cannot be compared to interventional effects because the models cannot produce such estimates. The data source determines what type and level of ground truth is available, constraining choices for evaluation measure, and an algorithm can only be evaluated using measures that reflect the model output. Different communities that work on causal modeling tend to focus on narrow pieces of this space. We begin by discussing some commonly used evaluation methods, then describe the components of an evaluation in more detail.

2.1 Data Sources

Data sources for evaluation can be be broadly divided into two categories: synthetic or empirical. We categorized data as empirical when it was collected from a "real world" system, whether that was a randomized clinical trial, a global financial system, or user interaction with a website. The important distinction is that empirical data was collected from a process or a system that exists for some purpose beyond scientific research.[1] Synthetic data includes anything else, including data generated from a randomly instantiated network structure or from a simulation intended to reflect a real-world system.

2.1.1 Synthetic Data

Researchers have developed several approaches to generating synthetic data. The most common is to use a directed graphical model or structural equation model [196, 68, 130]. Data can be generated from these models, and ground truth structure and effect estimates are readily available for evaluation. Other approaches involve designing the structure of a graphical model to match the causal structure of a realistic system. This can be done by manually specifying the structure based on domain knowledge [210, 93] or by learning a model from empirical data that can generate synthetic data [202]. Large-scale simulators designed for other reasons can also be used [16, 192]. In some cases, simulators can be complex enough to generate data that is effectively equivalent to empirical data, though such simulations vary widely in quality.

[1] Note that clinical trials are empirical under this definition. While the trial itself is performed for scientific research, the system under study was not constructed for the study (e.g., a human body reacting to a certain medication).

2.1.2 Empirical Data

Types of empirical data vary depending on the level of ground truth, and whether the ground truth came from a randomized controlled trial or prior knowledge of the domain. Purely observational data is the most readily available and is used most often. While this is rarely accompanied by full knowledge of the underlying structure, there are generally some dependencies that are known, either from common sense knowledge (such as temporal ordering) or from dependencies that have already been established by prior work [196, 68, 93]. For a randomized controlled trial, the dependence between the measured treatment and outcome is generally taken as ground truth [176]. The same is true for data where all potential outcomes are observed, where multiple different sets of interventions can be performed across the whole population or functionally identical individuals can be given different versions of treatment. This includes gene knockout studies, [130] flow cytometry analysis, [129] twin studies, [128] and computational systems [67].

There are a few data sets aimed at circumventing the typical limitations of real-world data. The 2016 Atlantic Causal Inference Conference (ACIC) Competition and subsequent competitions [53, 82] created semi-synthetic data sets. These data sets were created by producing synthetic treatment and outcome functions using covariates from a real-world system. Because the treatment and outcome functions are synthetic and known, true causal effects can be calculated. This approach is also used by the IBM Causal Inference Benchmarking Framework [174]. Another approach, described in more detail in Chapter 3, consists of collecting interventional data from a system where counterfactual intervention is possible. One example of this is large-scale computational systems; because they are run in a fully controlled computational environment, it is possible to apply a single treatment, collect observations, and then reset to the system to the pre-treatment state, allowing for the observation of all potential outcomes [64]. By doing this to all members of the pop-

ulation, true causal effects from treatment to outcome can be calculated. In both of these cases (ACIC competition data and the all-potential-outcomes data), it is possible to create *constructed observational data* (data that contains confounding, as is typical in observational data). This can be done by selecting treatment for each unit in a biased way based on observed covariates. Evaluation can then performed by running the algorithm on the observational data and comparing estimated causal effects with the known interventional effects.

2.2 Causal Modeling Algorithms

The algorithm under evaluation is not part of the evaluation method per se, but aspects of the algorithm strongly influence how evaluation can, and should, be performed. Algorithms can be broadly divided into two categories, *bivariate* and *multivariate*, although there are many variants. This distinction refers to more than just the number of variables the algorithm considers. Bivariate algorithms deal with a single causal dependence, one cause and one effect. While other variables may be considered, they are only included to improve estimation of the causal dependence between the two main variables. Multivariate algorithms, on the other hand, are focused on learning the dependence structure among a larger set of variables.

2.3 Evaluation Measures

At the heart of any evaluation technique is a measure of performance. Evaluation measures are generally designed to provide a single value that measures the 'correctness' of the learned causal dependencies. At a high level, evaluation measures fall into two categories, *structural* and *distributional*, based on whether they evaluate whether the structure or the parameters were learned correctly.

One other category, that can loosely be considered an evaluation, is what we refer to as *visual inspection*. This consists of looking at the learned structure or causal

effect estimates and, using domain knowledge, qualitatively describing if they look reasonable.

2.3.1 Structural Measures

Structural measures are designed to assess whether the structure (including both existence of edges and edge orientation) learned by the algorithm matches the ground truth. The most popular such measure for multivariate algorithms is structural Hamming distance (SHD) [191], which defines the distance between two graphs as the edit distance — the number of edge changes needed to make the graphs equivalent. In this setting, an edge change consists of adding, removing, or flipping the orientation of an edge. This is equivalent to counting the number of edges that were incorrect when compared to the ground truth. These measures are almost exclusively used with methods that produce causal graphical models.

Other related measures are used as well, such as precision, recall, F1-score, true-positive rate, and area under the ROC curve (AUROC) [36, 207, 15]. These are all closely related to SHD, since they are defined with respect to edge correctness (precision is defined as the fraction of learned edges that are correct, and recall is the fraction of correct edges that are learned). A variant of structural measures — structural intervention distance (SID) [153] — has been proposed though it is not used frequently in practice. Rather than counting the number of incorrect edges (which penalizes all edge errors equally), SID counts the number of interventional distributions that would be affected by the edge orientation errors.

2.3.2 Distributional Measures

Distributional measures are designed to capture how well the algorithm can estimate quantitative dependence. While structural measures can assess if the algorithm learned the correct model structure, a distributional measure is required to assess if the algorithm learned the correct parameters of that structure, and to assess the im-

pact of specific errors in structure learning. Such measures can be further subdivided into observational and interventional measures.

Observational measures compare the learned distribution with an observational ground truth (i.e., probability queries that do not involve a *do* operator). This could be classification error [182] or a measure of the error when predicting a given outcome variable [13, 66, 201]. Interventional measures, on the other hand, compare the learned distribution to ground truth obtained through intervention [176, 142, 195]. Measures of average and conditional average treatment effect (ATE and CATE, respectively) are common interventional measures [144], and KL-divergence and total variation distance [125] can be used for this purpose as well, when comparing estimated and ground truth interventional distributions.

While the actual measures used in observational and interventional cases may be similar or even identical, they are applied to different forms of ground truth. For example, KL-divergence can be used as either an observational or an interventional measure. The distinction is whether the estimated distribution is compared against an observational distribution or an interventional distribution.

Computing an interventional measure requires an estimate of the actual interventional effect. When that is the case, we can estimate the actual interventional distribution $P = P(Y = y|do(T = t))$ for any outcome y and treatment t. This known distribution can be compared to the estimated interventional distribution $\hat{P}$ from the causal model under evaluation. We then can use an interventional measure to compare the true interventional distributions P to the estimated distribution $\hat{P}$. One such measure is total variation distance [125].

Definition 2.3.1. Total Variation Distance

$$TV_{P,\hat{P},T=t}(Y) = \frac{1}{2} \sum_{y \in \Omega(Y)} \left| P\left(Y = y|do(T = t)\right) - \hat{P}\left(Y = y|do(T = t)\right) \right|,$$

where $\Omega(Y)$ is the domain of Y. For continuous distributions, TVD can be computed through an integral of differences in probability densities.

To summarize, there are many options available for evaluating causal modeling methods, ranging from simple structural comparisons in synthetic data to distributional comparisons in complex experimental data. However, the frequency with which these methods are used, and situations in which these methods are useful, vary significantly. The usage and utility of these methods is the focus of the next chapter.

CHAPTER 3

ASSESSMENT AND SUGGESTIONS FOR IMPROVEMENTS TO CURRENT PRACTICE

Evaluation is central to research in artificial intelligence and machine learning [44, 117]. How we evaluate algorithms determines our perception of the relative effectiveness and usefulness of different approaches, and this knowledge guides choices about future research directions. As Cohen and Howe explained three decades ago: "Ideally, evaluation should be a mechanism by which AI progresses both within and across individual research projects. It should be something we do as individuals to help our own research and, more importantly, on behalf of the field."[1]

As fields of science and engineering develop, protocols for evaluating key hypotheses in these fields should develop alongside them. In this chapter, we offer an empirical analysis of the set of techniques typically used to evaluate the accuracy of algorithms for learning causal models, and we show that this set could be substantially enhanced. The ultimate goal of most algorithms for causal modeling is to learn models capable of accurately estimating the effects of interventions in real-world systems. With this goal in mind, we would like to evaluate algorithms by comparing their estimates to measurements of actual interventional effects in a real-world system. In practice, though, many evaluations fall short of this ideal, most frequently using only structural or observational measures and synthetic data. Without the use of *empirical data*, our evaluations produce little information about whether our algorithms generalize

[1]Portions of this chapter appeared at NeurIPS 2019.

to real-world systems, and this greatly reduces their likelihood of widespread adoption by others outside of the field. Without the use of *interventional measures*, our evaluations produce little information about whether learned models will accurately estimate the effects of interventions, limiting their real-world utility.

Note that we do not argue for *replacing* the prevailing techniques for evaluation. These techniques have substantial value, both in assessing overall performance and in allowing fine-grained experiments to diagnose specific performance issues. Rather, we argue for *augmenting* the current suite of evaluation techniques to gather experimental evidence that the prevailing techniques cannot. We also do not contend that interventional measures and empirical data are entirely absent from current studies. A very small minority of recent studies use these techniques in combination. Rather, we argue that interventional measures and empirical data should be used routinely, and should be used in combination, for any serious study of algorithms for learning causal models. Indeed, the conclusions of most studies that lack such evaluation techniques should be considered exploratory and would benefit from additional evaluation.

3.1 Survey of Current Techniques

To assess how frequently different evaluation techniques are used in practice, we surveyed recent publications on causal modeling in computer science conferences. We collected papers from five recent UAI, NeurIPS, AAAI, ICML, and KDD conferences, as well as causality workshops held at UAI. We examined papers whose titles contained the terms 'cause', 'causal', or 'causality' and then narrowed this selection of papers to those that describe, propose, or evaluate a causal modeling algorithm. This resulted in a final set of 111 papers, of which 82% (91) reported any sort of evaluation.[2]

[2]When reporting survey results, we follow each percentage with a parenthesized number representing the raw count. The denominator for percentages is 91, except where otherwise noted.

The counts of papers included in the final survey are shown in Table 3.1. While some relevant papers may fall outside of our search parameters, this approach captures a reasonably representative sample of recent work in causal modeling, allowing us to infer which techniques are used in practice and how frequently these techniques are used.

Table 3.1. Recent causality papers included in survey

Venue	2014	2015	2016	2017	2018	**Total**
UAI	2	3	5	3	7	20
NeurIPS	3	5	4	6	13	31
AAAI	1	6	2	4	5	18
ICML	1	5	1	3	5	15
KDD	0	2	3	0	2	7
UAI-W	2	2	4	3	9	20
Total	9	23	19	19	41	111

The surveyed papers used a wide range of data sources, but we broadly categorize them as empirical or synthetic. In our survey, we found many examples of both, and while synthetic data is used more frequently, both are still common. 81% (74) of papers surveyed used synthetic data, 67% (61) used empirical data, and 48% (44) used both.

We also analyzed the categories of algorithms used by papers in the survey. Multivariate algorithms are significantly more prevalent in the data, accounting for 60% (55/111) of papers surveyed. Bivariate algorithms account for 30% (34/111) of papers surveyed, split between those focused on orientation (10%), magnitude of effect (15%), or both (5%). The remaining papers in the survey fall in between, including those that aim to determine the joint effect of multiple treatment variables on a single outcome.

We also analyzed the evaluation measures used. The main types we consider are structural measures (such as structural hamming distance, precision/recall), obser-

vational measures (such as RMSE and classification accuracy), and interventional measures (such as total variation distance, average treatment effect).

Of the three types of evaluation measures, structural measures are the most common, being used in 55% (50) of papers surveyed. Distributional measures are slightly less common, being used in 46% (42) of papers. The vast majority of the distributional measures used, however, are observational rather than interventional; observational measures are used in 33% (30) of papers, while interventional measures are used in only 13% (12).

The choice of evaluation measure depends on both the data generation process and type of algorithm, which is reflected in our survey. When synthetic data is evaluated, structural measures are used 59% (44/74) of the time. However, when empirical data is evaluated, structural measures are used only 38% (23/61) of the time, since empirical data is less likely to have ground truth. This lack of ground truth sometimes prevents any significant evaluation for techniques using empirical data—26% (16/61) of empirical evaluations consisted exclusively of visual inspection of the results, with no ground truth. Table 3.2 summarizes the interaction between data source and evaluation measure in the survey.

Table 3.2. Summary of survey results: the number of papers using different evaluation measures and data sources

		Data Sources	
		Synthetic	Empirical
Evaluation Measures	Visual Inspection	0	19
	Structural	44	23
	Observational	22	15
	Interventional	11	5

The survey makes clear that the vast majority of papers that perform evaluations use either: (1) synthetic data; or (2) empirical data combined with non-interventional measures (observational measures, structural measures, or visual inspection). Our

proposed ideal evaluation (empirical data and interventional measures) is used in only 5% (5) of papers. This raises an obvious question: Are the most commonly-used evaluation techniques sufficient for determining whether algorithms for learning causal models will work effectively in realistic scenarios? As we will argue below, they are not.

3.2 The Case for Empirical Data

As already noted, nearly all causal modeling algorithms are ultimately designed for use outside of a laboratory—on real systems to infer useful causal knowledge about the world. Despite this, evaluation of such algorithms often uses synthetic rather than empirical data.

3.2.1 Limitations of Synthetic Data

Researchers have developed several approaches to generating synthetic data. The most common is to use a directed graphical model or structural equation model. Other approaches involve designing the structure of a graphical model to match the causal structure of a realistic system, either by manually specifying the structure or by learning it from empirical data. Large-scale simulators designed for other reasons can also be used. In some cases, simulators can be complex enough to generate data that is effectively equivalent to empirical data, though such simulations vary widely in quality.

Synthetic data is easy to collect, allows for straightforward comparison with ground truth, and facilitates systematic testing across a variety of data parameters. Its popularity is evident—84% (74) of surveyed papers used it in their evaluation, and 41% (30/74) of those used only synthetic data. However, using synthetic data for evaluation also has significant limitations. These include:

Unquestioned assumptions—Synthetic data tends to match the assumptions of the researcher running the study and any algorithms they have created. For example, a researcher developing an algorithm that outputs a DAG will be inclined to generate data from a DAG.

Untested influences—Even the best data generators can only include the influences already known to researchers. Almost by definition, synthetic data generators cannot include any "unknown unknowns" that may influence the outputs of real-world systems.

Lack of standardization—Synthetic data is typically generated differently by each researcher, and this lack of standardization impedes comparison between studies.

Researcher degrees-of-freedom—Synthetic data is typically designed and parameterized by the researchers who created the algorithm being evaluated, giving them an enormous range of choices. Such high "researcher degrees-of-freedom" [177] are a basic challenge to the validity of any study.

These factors significantly limit the external validity and realism of most synthetic data, making it insufficient as the sole source of data for evaluation. Synthetic data is not without value—it can be a powerful way to assess features of an algorithm and test its performance under different conditions. However, it typically falls short in providing insights into how the algorithm will perform on data from a real-world system.

3.2.2 Sources of Empirical Data

Types of empirical data vary depending on the level of ground truth, and whether the ground truth came from a randomized controlled trial, an interventional experiment, or prior knowledge of the domain. Purely observational data is the most readily available and is used most often. While this is rarely accompanied by full knowledge of the underlying structure, there are generally some dependencies that are known, ei-

ther from common sense knowledge (such as temporal ordering) or from dependencies that have already been established by prior work. For a randomized controlled trial, the dependence between the measured treatment and outcome is generally taken as ground truth. The same is true for interventional experiments, which we define to be experiments performed on a system where multiple different sets of interventions can be performed across the whole population. This includes gene regulatory networks, flow cytometry analysis and computational systems, where the system can be rerun with new parameter settings.

3.2.3 Benefits of Empirical Data

Empirical data is almost always more difficult to collect than simulated data, and information on the effects of interventions is typically also more difficult to obtain. However, using empirical data has multiple benefits:

Realistic complexity—Empirical data typically has a distribution that is more complex than simulated data. That distribution is subject to realistic latent factors and measurement error. This creates a learning task that is often significantly harder than synthetic data, but also more closely matches the challenges of real-world settings.

Lower potential researcher bias—Empirical data is typically not generated by the researcher who designed the algorithm being evaluated, and thus it is less subject to unintentional biases. In addition, individual data sets are often shared across the community, creating standardization and comparability across studies.

Real-world demonstration—The aim of research on algorithms for causal modeling is to have these algorithms used by others to infer causal models and reason about causal effects in real-world settings. Practitioners considering use of these methods may be legitimately skeptical about their effectiveness until they see successful demonstrations of accurate causal modeling on real-world data.

However, using empirical data poses challenges as well. Because it is generally not collected by the person using it, there may be features of the data that are not fully understood, hindering correct interpretation. Also, ground truth can be challenging to obtain, limiting evaluation to visual inspection or observational measures. This is unsatisfying at best and misleading at worst, since, when evaluating without ground truth, it can be easy to see meaning where none exists or to imagine explanations for many possible conflicting outputs. Despite these challenges, empirical data, while less common than synthetic, is still used frequently in practice; 67% (61) of surveyed papers use empirical data, and 28% (17/61) used only empirical data.

3.2.4 Are Empirical Data Sets Available?

Because interventional measures and empirical data are used so infrequently, one may assume this is because such data sets are difficult to find. This is partially true—there are significantly more observational data sets available than interventional data sets. However, there is a growing community that is producing data sets that provide interventional effects. We describe some of them here.

The cause-effect pairs challenge [143] provides a data set that is empirical and, while interventional effects are not available, the direction of causality is known. The 2016 Atlantic Causal Inference Conference (ACIC) Competition and subsequent competitions [53, 82], created semi-synthetic data sets, producing synthetic treatment and outcome functions using covariates from a real-world system. A similar approach was used for the IBM Causal Inference Benchmarking Framework [174]. Gene regulatory networks, specifically the DREAM in silico data sets, are a popular choice, since multiple combinations of single-gene interventions can be performed on identical networks [165]. The DREAM data sets are taken from a sophisticated simulation derived from multiple known gene regulatory network structures, which, while non-empirical, is intended to be complex enough to approximate empirical data. Flow cytometry data,

measuring protein signaling pathways, is another common choice for interventional data, specifically the data set provided by Sachs et al. [164]. Flow cytometry allows for the simultaneous measurement of large signaling networks, and many independent cells can be measured, allowing for the analysis of multiple interventions. Dixit et al. [52] provide data on gene expression, collected using their proposed Perturb-Seq technique to perform gene deletion interventions. Other sources of interventional and empirical data include results of advertising campaigns [184] and clinical studies [136], as well as multiple challenges organized for machine learning conferences [80, 81].

Garant and Jensen [64] introduced an additional source of empirical data where interventions are possible: large-scale computational systems. They performed experiments on three large computational systems: Postgres, the Java Development Kit, and HTTP processing. These systems have many desirable properties for the purposes of empirical evaluation: (1) They are pre-existing systems created by people other than the researchers for a purpose other than evaluating algorithms for causal modeling; (2) They produce non-deterministic experimental results due to latent variables and natural stochasticity; (3) System parameters provide natural treatment variables; and (4) Each experiment is recoverable, allowing the same experiment to be performed multiple times with different combinations of interventions.

Within each computational system, three classes of variables are measured: outcomes, treatments, and subject covariates. Here, outcomes are measurements of the result of a computational process, treatments correspond to system configurations and are selected such that they could plausibly induce changes in outcomes, and subject covariates logically exist prior to treatment and are invariant with respect to treatment. Using these variables, all combinations of treatments can be applied to all subjects, and we can use these results to estimate actual interventional distributions for the effects of each treatment variable on each outcome variable. We can also then sub-sample these experimental data sets in a manner which simulates observational

bias to produce observational-style data sets, allowing us to evaluate an algorithm's performance on pseudo-observational data and evaluate it using actual interventional effects. Many additional details on the generation and use of these data sets are presented in Appendix B.

3.2.5 How Different are the Results?

Readers may ask: In practice, what's the difference between using empirical data rather than synthetic data? If that difference is small, then the substantial extra work involved in evaluation with empirical data may not be worth the effort.

To begin addressing this question, we conducted a series of experiments using the interventional data from the computational systems described above. Specifically, we used a common approach for generating somewhat realistic synthetic data. This approach uses an empirical data set to learn a causal model and then uses that model to generate synthetic data (and known ground truth) for model evaluation. While the final data set is synthetic, its structure may better approximate the empirical system, rather than being entirely defined by the researcher, lending it more credibility. We used this approach to generate synthetic data in the style of the three empirical data sets we generated from computational systems. Since we now have both empirical and synthetic data, each with ground truth, we can use causal modeling algorithms to construct a model for both of these data sets and compare the conclusions we would draw from each.

The synthetic data used was created by first choosing an initial causal modeling algorithm to create a ground truth network from the empirical data. After learning a ground truth model with each of two algorithms that construct causal graphical models (PC and GES), we generated synthetic data using the resulting models. We then evaluated the same three algorithms on both the synthetic and empirical data. Figure 3.1 shows how mean TVD varies for different causal modeling algorithms and

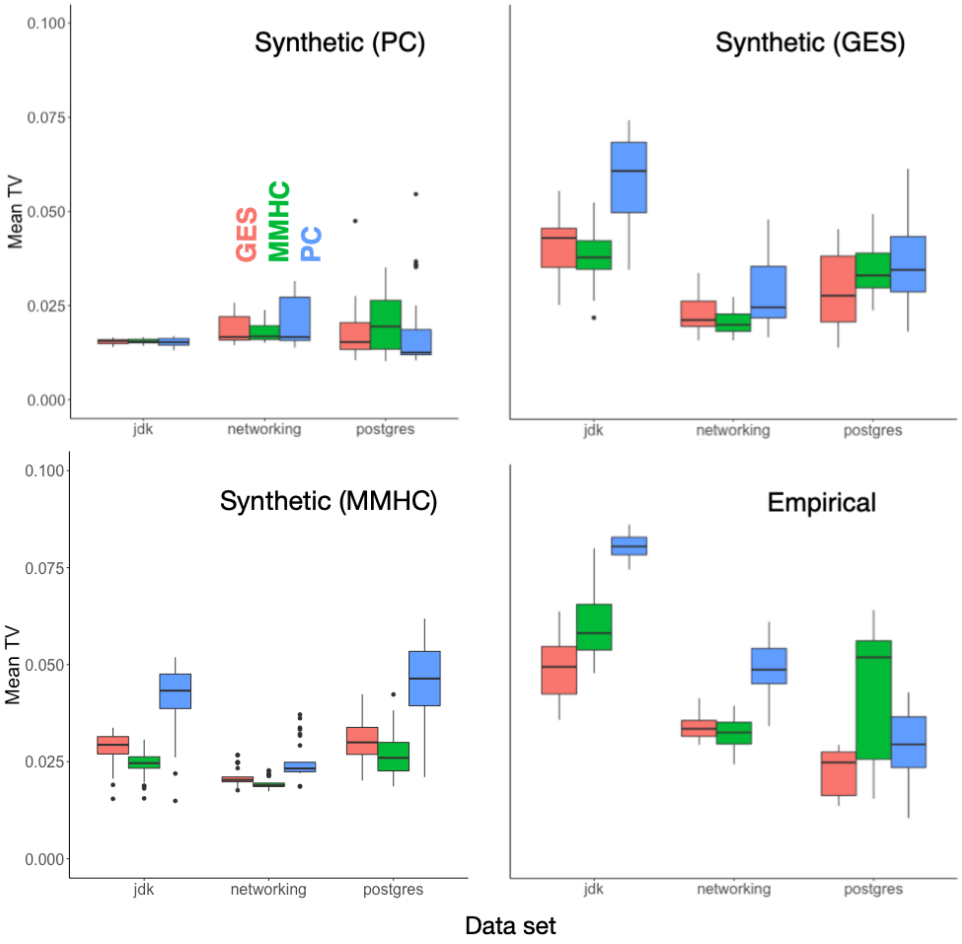

Figure 3.1. Comparison of TVD on empirical data and synthetic data derived from empirical data. *top and bottom left*: synthetic data, structure obtained from using PC (top left), GES (top right), or MMHC (bottom left). *bottom right*: TVD on empirical data.

different data sets. Because sample sizes for some of the computational system data sets are small, results are reported as distributions over 30 trials for each algorithm and data set. The results shown are the mean TVD when evaluating PC, GES, and MMHC on two types of synthetic data sets (using the model as ground truth) and on the empirical data (using the known interventional effects). There is significant variability between the two methods of generating the synthetic ground truth network from the empirical data (PC and GES), both in the mean TVD and in the relative ordering of the algorithms. Comparing the synthetic and empirical results, some relative orderings of the algorithms are the same (e.g., network), but other orderings are significantly different (e.g., Postgres). These results suggest that algorithm performance cannot be expected to match between synthetic and empirical data, even when the synthetic data is created in a way that would be most expected to match aspects of the empirical data.

3.3 The Case for Interventional Measures

Many algorithms are currently evaluated based on their ability to learn causal structure. However, the actual desired underlying task is almost never to model structure alone. In practice, estimating the magnitude of interventional effects is vitally important, and an algorithm that cannot distinguish between strong and weak effects is severely limited in scope. Despite this, the majority of current evaluations use observational or structural measures rather than measures of interventional effect.

3.3.1 Limitations of Observational Measures

Observational measures are generally used when the task of the algorithm is to discern the statistical association between two or more variables. They are generally applied in cases when the structure is not the primary focus or is already known, and the primary concern is effectively modeling the magnitude and form of dependence,

rather than the existence of dependence. However, observational data has a severe and obvious limitation:

Non-causal—Observational measures are, by definition, not causal. They measure the error of estimates of the outcome variable, but they do not measure that error under intervention. They provide a sense of how well an algorithm has learned statistical dependence, but not how well it has learned causal dependence. Despite this, observational measures are the only evaluation used in 24% (22/91) of papers surveyed.

3.3.2 Limitations of Structural Measures

Structural measures are easy to calculate, and they have a clear intuition. If an algorithm produces a causal structure and we know structural ground truth, it seems sensible to determine if the two structures match. This has led to the widespread adoption of structural measures: 55% (50) of surveyed papers used such measures, and 84% (42/50) of those used only structural measures. However, structural measures have several serious limitations:

Requires known structure—Calculating structural measures requires a full ground truth network structure, which is only rarely available for empirical data.

Constrains research directions—The prevalence of structural measures may constrain research to algorithms that can be evaluated with these measures. Algorithms that do not produce DAGs are less likely to be developed or favorably reviewed. Structural measures also implicitly assume that DAGs are capable of accurately representing any causal process being modeled, an unlikely assumption.

Oblivious to magnitude of dependence—Structural measures, by design, do not account for different magnitudes of dependence, so an error in an edge with a strong effect is weighted the same as an error in an edge with a very weak effect.

Oblivious to likely treatments and outcomes—In most cases, structural measures do not consider where an edge is located in the overall structure of the network, so an edge with many downstream effects is treated the same as a less central edge.

3.3.3 Benefits of Interventional Measures

In contrast to observational and structural measures, interventional measures have multiple advantages:

Correspondence to actual use—Interventional measures evaluate how well the model estimates interventional effects, which aligns more accurately with the eventual use of nearly all causal models. For example, a directed acyclic graph is not the ultimate artifact of interest for most applications; DAGs are a representation that facilitates estimation of interventional effects [151, 179]. Thus, it seems natural to define an evaluation measure in terms of interventional effects rather than graphical structure.

Weighting of different errors—While most structural measures weight each edge misorientation equally, interventional measures penalize misorientation errors proportionally to their effect on the estimation of interventional effect.

3.3.4 How Different are the Results?

Interventional measures are intended to capture something different than structural measures, but they are ultimately affected by the structure of the learned model, and we would expect structural errors to lead to interventional errors. To assess the extent to which interventional measures capture different information than structural measures, we ran experiments using synthetic data. This allowed us to produce data where we could calculate both structural measures and interventional measures, since we have the full parameterized ground truth model to compare against.

For these experiments, we produced data from random DAG structures with conditional probability models drawn from a Dirichlet distribution. We generated 5000

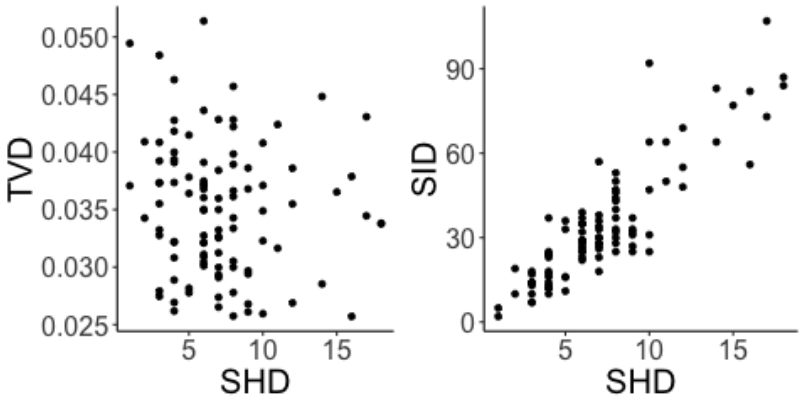

Figure 3.2. Structural and interventional measures compared on synthetic data with GES. *Left*: SHD And TVD, *Right*: SHD And SID

instances, applied a causal modeling algorithm, and calculated various evaluation measures. Figures 3.2, 3.3, and 3.4 shows the results for GES, MMHC, and PC, respectively. Interestingly, while the correlation between SID and SHD is relatively consistent for all three structure learning algorithms, the correlation between TVD and SHD varies substantially, from seemingly completely uncorrelated (GES) to very clearly correlated (PC). The strong correlation between SHD and SID suggests that both these structural measures ultimately produce similar quality measures of the algorithm. When comparing SHD and TVD, for some cases, such as GES, they are only very weakly correlated, with many models scoring highly with one measure and poorly with the other. However, for other cases, such as PC, structural measures appear to provide a decent proxy for interventional measures. However, it is unlikely that the researcher knows this to be the case ahead of time, and the comparative difference in TVD between the three algorithms suggests the value of using TVD when comparing multiple causal learning algorithms.

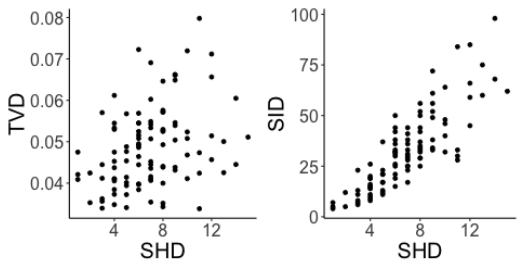

Figure 3.3. Structural and interventional measures compared on synthetic data with MMHC.

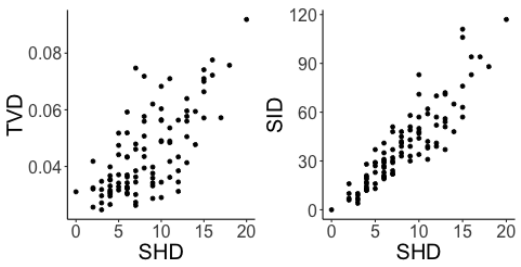

Figure 3.4. Structural and interventional measures compared on synthetic data with PC.

3.4 Example of an Evaluation

To further explain what we mean by empirical data and interventional measures, we describe one example of this type of evaluation, shown schematically in Figure 3.5. This example demonstrates one way that an evaluation with empirical data and interventional measures could be performed, though many other techniques are possible, depending on the algorithm, data source, evaluation measure, and the research question under consideration. In our example, we evaluate the PC algorithm [179], Greedy Equivalence Search (GES) [41], and MMHC [191] by measuring *total variation distance* (an interventional measure defined later) on a data set produced by experimentation with a large-scale computational system.

An obvious way to evaluate how well an algorithm can learn causal models from real-world data is to compare the model's estimates to empirical data drawn from a system in which we can perform multiple interventions on the same units, giving us full interventional data in which we can assess every potential outcome for each unit. Large-scale computational systems allow for this type of intervention because they let us run the same experiments multiple times under different conditions (e.g., different settings of key system parameters). An example of this is a Postgres database, where we can run the same queries with different settings of key configuration parameters. In this context, each query corresponds to a unit, a set of configuration parameters

corresponds to a treatment, and variables such as runtime correspond to outcomes. Details about these data sets can be found in Appendix B.

Many algorithms for causal modeling are designed to run on observational data, in which only a single, non-randomized treatment assignment is observed for each unit. In the absence of an observational data set that matches our interventional data, we can create an observational-style data set by sub-sampling the full interventional data in a non-random manner. To do this, we select a single treatment assignment for each query. Selecting treatment at random is equivalent to a randomized controlled trial. In most observational contexts, however, treatment assignment would be based on covariates of the units. For example, a database administrator might choose the configuration parameters based on features of each query. We use a similar process to create observational data by using a measured covariate of the query to probabilistically assign treatment.

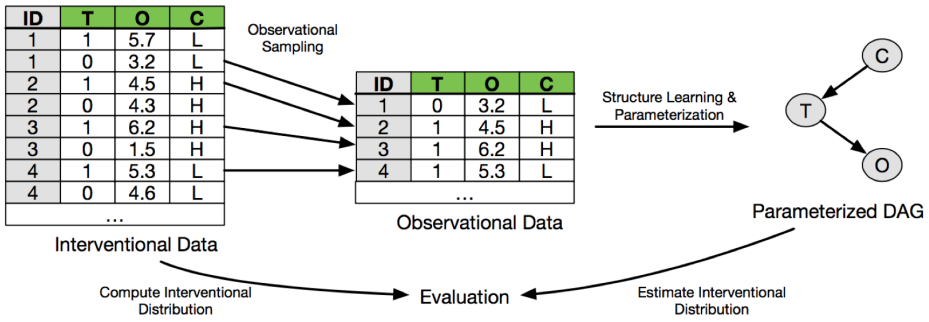

Figure 3.5. A diagram of one approach to evaluating a causal modeling algorithm

Given such an observational data set, we can apply a causal modeling algorithm and learn a causal model. A fully parameterized model can produce an estimated interventional distribution $\hat{P}$ by applying the *do*-calculus [61]. Under this framework, causal quantities take the form of probability queries with *do* operators, for instance $P(O|do(T = 1))$. We can also estimate the actual interventional distribution $P =$

$P(O = o|do(T = t))$ for any outcome o and treatment t, because we can measure the effects of both values of treatment for each query in our data set.

We then can use an interventional measure to compare the true interventional distribution P to the estimated distribution $\hat{P}$. One example of an interventional measure is total variation distance (TVD) [125], which measures the distance between two probability distributions. For discrete outcomes O, the quality of an estimated interventional distribution relative to a known distribution under TVD can be computed as described in Equation 4.1. This gives us a numerical measure of how well the estimated interventional estimates match the ground truth. A single TVD value is computed for each causal effect, which can then be aggregated for comparison. Results of this evaluation on the computational data is shown in Figure 3.1c. For these data sets, we can conclude that GES has the best overall performance.

3.5 Conclusions

Evaluation is a key mechanism that determines how algorithms are viewed within the community, what research directions are pursued next, and whether our research has broader impacts outside the community. Our current evaluation techniques aim too low, and they fail to evaluate the full range of questions that our research goals imply.

We acknowledge that, while the evaluation techniques we advocate are applicable to wide range of algorithms, data sets may not be available for every task. The diverse tasks of causal modeling algorithms make it difficult to recommend a single data set and evaluation measure to evaluate every algorithm. However, the data sets and measures that are most commonly used are largely insufficient. We believe it would benefit the community for more data sets with interventional effects to be created and made available for public use, allowing for a breadth of evaluation options.

We do not advocate abandoning synthetic data and structural measures. Both have many uses for evaluating algorithm performance and can be indispensable scientific tools. However, they are not sufficient on their own. Instead, they should be viewed as a first step in evaluation. If novel algorithms for causal modeling are to be widely adopted, prospective users will justifiably require credible demonstrations of their utility outside of a laboratory setting. If we do not evaluate on empirical data, we cannot be certain our algorithms will perform well on real data, and if we do not evaluate with interventional measures, we cannot be certain that the causal effects the algorithm infers will translate to actual, substantial causal effects. Expanding our routine evaluations will substantially improve the credibility and comparability of results, the external validity and trustworthiness of algorithms, and the efficiency with which we conduct our research.

CHAPTER 4

USING EXPERIMENTAL DATA TO EVALUATE METHODS FOR OBSERVATIONAL CAUSAL MODELING

4.1 Introduction

Most easily 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 modeling 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 [53, 67, 192, 174], 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 repurpose 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 and thus limiting research progress.

We propose to exploit an additional source of data for evaluating causal modeling methods: randomized controlled trials. Randomized controlled trials (RCTs) are designed and conducted for the express purpose of providing unbiased estimates of treatment effect. Many RCT data sets are publicly available, and more become available every day. Previous work has described how to sub-sample a specialized type of experimental data (one in which all potential outcomes are observed) to create

constructed observational data.[1] Surprisingly, this basic approach can be modified to produce constructed observational data from RCTs as well. Specifically, we: (1) Describe an algorithm to induce confounding bias in RCT data by sub-sampling, and 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 modeling; (2) Demonstrate the feasibility of this approach by applying multiple causal modeling methods to observational data constructed from RCTs;[2] and (3) Present a method for using the data rejected by the sub-sampling for evaluation, and show that it is equivalent to a held-out test set.

4.2 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.

---

[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 "Related Work").

[2]Pointers to the data sets used in this paper, and R code to perform observational sampling, will be provided at `http://kdl.cs.umass.edu/data`

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

In Chapter 3, we proposed sampling from APO data to produce constructed observational data. A small number of other researchers (e.g., [128]) have proposed similar procedures. 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 ($P(T|C) := f(C)$). We refer to the set C as the *biasing covariates*. This procedure, shown in Algorithm 1, induces causal dependence between C and T, creating a confounder when an element of C 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 [163].

Data sets produced by OSAPO are extremely useful for evaluating causal modeling methods. Causal modeling methods can estimate treatment effect from 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 are observed. However, very few experimental data sets exist that record all potential outcomes for every unit, severely limiting the applicability of this approach.

4.2.1 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 implemented by RCTs, in which every unit is randomly assigned a 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?

Algorithm 1: Observational sampling from all potential outcomes (OSAPO)

Input : APO data set D_{APO}, biasing covariates C
Output : Biased data set D_{OSAPO}
foreach *unit* $i \in D$ **do**
  $p \leftarrow f(c_i)$
  $t_s \leftarrow Bernoulli(p)$
  $o \leftarrow$ row in D_{APO} corresponding to (i, t_s)
  $D_{OSAPO} \leftarrow D_{OSAPO} \cup o$
end
return D_{OSAPO}

Algorithm 2: Observational sampling from randomized controlled trials (OSRCT)

Input : RCT data set D_{RCT}, biasing covariates C
Output : Biased data set D_{OSRCT}
foreach *unit* $i \in D$ **do**
  $p \leftarrow f(c_i)$
  $t_s \leftarrow Bernoulli(p)$
  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
end
return D_{OSRCT}

Figure 4.1. 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 $f : \mathcal{D}(C) \rightarrow \{p \in \mathbb{R} : 0 < p < 1\}$

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 bias the selection of a single treatment value 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, when treatment is binary and 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 4.2.2, a causal modeling 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 4.2.

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 OSAPO. 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}.

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.

Proof. For every unit i and any treatment t', the biasing covariates C_i are used to probabilistically select a treatment, which we denote T_{si}, with probability $P(T_{si} = t'|C_i)$.

$$P_{D_{OSAPO}}(T_i = t') = P(T_{si} = t'|C_i) \tag{4.1}$$

$$P_{D_{OSRCT}}(T_i = t') = P(T_i = t')P(T_{si} = t'|C_i) \tag{4.2}$$

$$= 0.5P(T_{si} = t'|C_i) \tag{4.3}$$

Sub-sampling D_{OSAPO} uniformly at random is equivalent to multiplying $P_{D_{OSAPO}}(T_i = t')$ by a scaling factor, s. When $s = 0.5$, $P_{D_{OSRCT}}(T_i = t') = P_{D_{OSAPO}}(T_i = t')$. □

Intuitively, the procedure outlined in Algorithm 4.1 works because treatment is *randomly* assigned in RCTs. The data is sub-sampled based solely on the value of a probabilistic function of the biasing covariates, which selects a value of treatment for every unit i. Since the observed treatment is randomly assigned, it contains no information about any of i's covariates. The only bias introduced by this sub-sampling procedure is the intended bias: a particular form of causal dependence from C to 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 $E[P(T_s = 1|C)] = 0.5$,
then $E[|D_{OSRCT}|] = 0.5|D_{RCT}|$.

Proof. Assume binary treatment $T \in \{0, 1\}$. For any unit i with covariates C_i, let $P(T_i = t) = p_t$, $P(T_{si} = t|C_i) = p_{T_{si}=t|c}$, and $n = |D_{RCT}|$. Indices are omitted when clear from context.

$$\begin{aligned}
P(i \in D_{OSRCT}) &= p_1 p_{T_{si}=1|c} + p_0 p_{T_{si}=0|c} \\
&= p_1 p_{T_{si}=1|c} + (1 - p_1)(1 - p_{T_{si}=1|c}) \\
&= 2p_1 p_{T_{si}=1|c} - p_1 - p_{T_{si}=1|c} + 1 \\
E[|D_{OSRCT}|] &= \sum_{i=1}^{n} [2p_1 p_{T_{si}=1|c} - p_1 - p_{T_{si}=1|c} + 1] \\
&= n - np_1 + (2p_1 - 1) \sum_{i=1}^{n} p_{T_{si}=1|c}
\end{aligned}$$

If either $p_1 = 0.5$ or $\sum_{i=1}^{n} p_{T_{si}=1|c} = 0.5n$, $E[|D_{OSRCT}|] = 0.5n$. □

4.2.2 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 modeling methods. In a well-designed RCT, treatment assignment is randomized

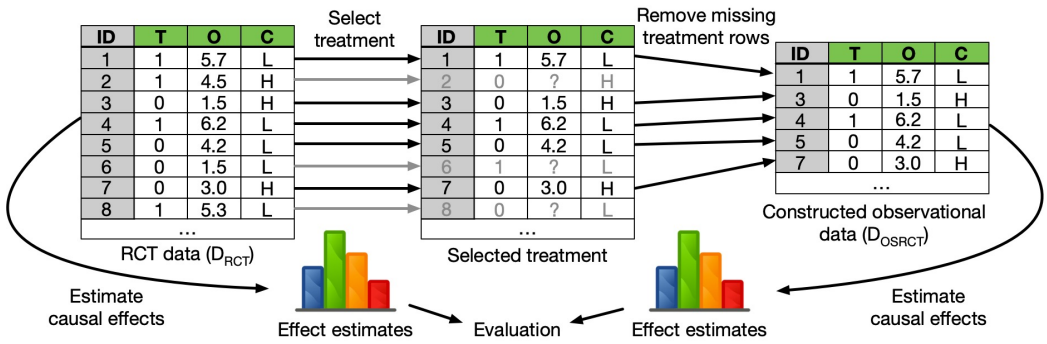

Figure 4.2. The process of creating observational-style data from a randomized controlled trial.

such that, in expectation, the treatment and control groups are equivalent. This enables the unbiased estimation of the sample average treatment effect (ATE) as $E[y_i(1)|t_i = 1] - E[y_i(0)|t_i = 0]$, where t_i denotes the actual treatment received by unit i. This estimate can be compared to estimates made by causal modeling methods applied to the constructed observational data.

However, 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 measurement which *is* available when using APO data. However, OSRCT data *can* be used to evaluate a method's ability to estimate the unit-level effects of interventions. Any causal modeling method that can estimate $E[Y|do(T = t)]$ can be evaluated by comparing those estimates against measurements in the RCT data.

4.2.3 Using the Complementary Sample for Evaluation

One challenge when evaluating causal modeling methods on their ability to estimate unit-level effects of interventions is the need for a held-out test set. The constructed observational data is constructed by sub-sampling the original RCT data.

This means that evaluating on all of the RCT data may produce a biased result by testing on a superset of the training data. One potential solution is to divide the RCT data into separate training and test sets. However, since OSRCT necessarily reduces the size of the training data by sub-sampling, the extra requirement of holding out a test set limits the number of RCTs that can be used, since not all randomized experiments will have enough data to learn effective models after two rounds of sub-sampling.

A more data-efficient approach is to use the data rejected by the biased sub-sampling. OSRCT sub-samples RCT data to create a probabilistic dependence between the biasing covariates and treatment. Based on the values of the biasing covariates, a treatment is selected for every unit. If that treatment is present in the data, the unit is included in the sample; otherwise the unit is rejected. This rejected sample (which we call the *complementary sample*) also has a causal dependence from the biasing covariates to treatment. The only difference is that the form of that dependence is the complement of that for the accepted sample, such that covariate values that lead to a high probability of treatment in the accepted sample would lead to a low probability of treatment in the complementary sample. Because we know the functional form of this induced bias, 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 [158].

Theorem 3. *For binary treatment $T \in \{0, 1\}$, biasing covariates C, outcome Y, estimated outcome $\hat{Y}$, biased sample D_{OSRCT} and complementary sample $\bar{D}_{OSRCT}$, let $p_s = P(T_{si} = t_i | C_i)$. Then, $E[\hat{Y} - Y]$ for $D_{OSRCT} = E[(\hat{Y} - Y)\frac{p_s}{1-p_s}]$ for $\bar{D}_{OSRCT}$.*

Proof. For D_{OSRCT},

$$E[\hat{Y} - Y]_{D_{OSRCT}} = E[P(T_{si} = t_i|C_i)(\hat{Y}_i - Y_i)]$$

For $\bar{D}_{OSRCT}$,

$$E[\hat{Y} - Y]_{\bar{D}_{OSRCT}} = E[(1 - P(T_{si} = t_i|C_i))(\hat{Y}_i - Y_i)]$$

If we weight the outcome estimates for $\bar{D}_{OSRCT}$ by $\frac{P(T_{si}=t_i|C_i)}{1-P(T_{si}=t_i|C_i)}$,

$$\begin{aligned} E[\hat{Y} - Y]_{\bar{D}_{OSRCT}} &= E[\frac{P(T_{si} = t|C_i)}{1 - P(T_{si} = t_i|C_i)} \cdot \\ &\quad (1 - P(T_{si} = t_i|C_i))(\hat{Y}_i - Y_i)] \\ &= E[P(T_{si} = t_i|C_i)(\hat{Y}_i - Y_i)] \\ &= E[\hat{Y} - Y]_{D_{OSRCT}} \end{aligned}$$

□

4.2.4 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 modeling.

Evaluation with OSRCT has some limitations. OSRCT can induce dependence between any covariate and treatment, but it cannot induce dependence 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 factors of observational studies can also be simulated, such as measurement error, selection bias, and lack of positivity. While some of these may reduce the sample size of the constructed observational data due to additional sub-sampling, this can allow for the evaluation of a causal modeling method's robustness to many features of real-world data.

4.3 Related Work

We deferred discussion of some related work because it directly and exclusively relates to the content of this chapter. We discuss that work below. The closest prior work [124] uses an identical idea for a subtly different task: estimating the reward of a contextual bandit policy without having to actually execute that policy. Specifically, they propose to evaluate a (non-random) contextual policy by sampling from the data produced by a randomized policy. They show that the resulting estimate is unbiased, despite its use of only a subsample of the data originally produced by the randomized policy. This method is widely employed to evaluate methods in fields such as computational advertising and recommender systems, and it has been extended with approaches such as bootstrapping [135].

OSRCT exploits the same idea but in a different setting. In our setting, we have no interest in estimating the effect of a contextual policy that is known to the agent (which is somewhat analogous to what, in observational causal modeling, is referred to as the "average treatment effect on the treated"). Instead, our goal is to determine how well a given method estimates the average 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.

Despite the similarity of tasks, this approach—observational sampling from RCTs—is almost entirely unknown within the causal modeling community. For example, two recent papers that contain reviews of existing evaluation methods for causal modeling methods—Dorie et al. [53] and our own recent work [67]—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). A handful of papers have applied it in a one-off manner to evaluate causal modeling methods [103, 104], but it has not been explicitly formalized or its advantages clearly described. As a result, it is almost never used.

In addition to this prior work on sampling for evaluating contextual bandit policies, other prior work has explicitly focused on evaluation methods in causal modeling. This work has applied a variety of approaches to creating observational data sets such that a derived estimated treatment effect can be compared to some objective standard. The ideal approach would score highly on at least three characteristics: *data availability* (many data sets with the required characteristics can be easily obtained); *internal validity* (differences between estimated treatment effect and the standard can only be attributed to bias in the estimator); and *external validity* (the performance of the estimator will generalize well to other settings). Of three broad classes of prior work, each suffers from some deficiencies and none clearly dominate the others.

The first class of 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 [143]. Unfortunately, such situations are relatively rare, limiting data availability. Another approach is to use data from

matched pairs of observational and experimental studies [52, 164]. In many ways, such data sets appear to represent a nearly ideal scenario for evaluating methods for inferring causal effect from observational data. However, 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 [116, 89, 168], but these studies also have low data availability.

Another class of prior work generates observational data from *synthetic or highly controlled causal systems* [192, 67, 128, 101]. 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 [53, 174]. Given the synthetic nature of the outcome, 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, these 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 be applied to data from any moderately large RCT. Only synthetic data generators and approaches that augment observational data with synthetic outcomes probably 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.

4.4 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 modeling 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. This includes repositories such as Dryad, the Yale Institution for Social and Policy Studies Repository, the NIH

National Institute on Drug Abuse Data Share Website, the University of Michigan's ICPSR repository, the UK Data Service, and the Knowledge Network for Biocomplexity. An even larger set of repositories restricts access but will make data available upon reasonable request.

In addition, funding agencies and journals are increasingly requiring that researchers make anonymized individual patient data available upon reasonable request [73, 149]. 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 [1]. 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 [54, 114, 24, 187]. All of this suggests increasing availability of individual patient data from randomized controlled trials.

4.5 Experimental Evaluation

To assess OSRCT's effectiveness at approximating APO data, we performed an experiment using the Postgres data discussed in Appendix B. In this data, units are Postgres queries, interventions are Postgres settings (such as type of indexing), covariates are features of queries (such as the number of joins or the number of rows returned), and outcomes are measured results of running the query (such as runtime). If the Postgres database is queried in a recoverable manner, the same query can be run repeatedly while varying the treatment, creating APO data. For this analysis, we chose runtime as the outcome, indexing level as the treatment, and the number of rows returned by the query as the biasing covariate.

To compare RCT and APO data, We converted the APO Postgres data into RCT-style data by randomly sampling a single treatment for every unit. We then created constructed observational data from both the original APO data and the RCT-style

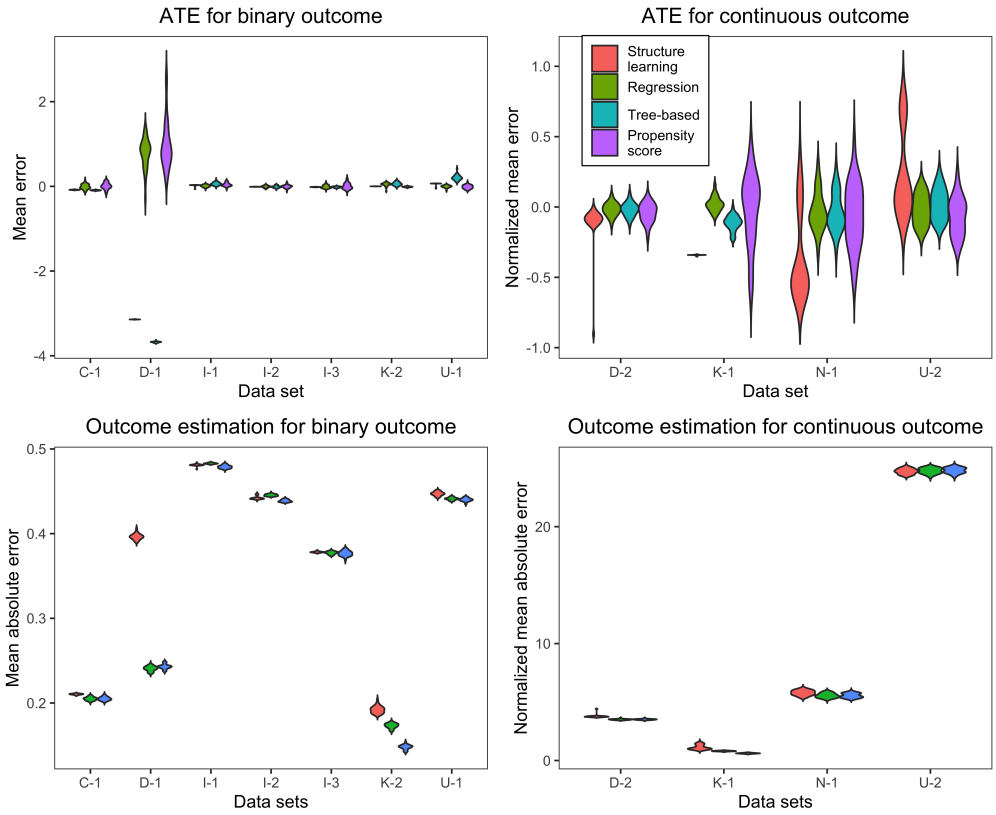

Figure 4.3. Demonstration of OSRCT on data from 11 RCTs, split by outcome type. *Top left:* ATE (risk-ratio) for binary outcome, *Top right:* ATE for continuous outcome, *Bottom left:* Outcome estimation for binary outcome, *Bottom right:* Outcome estimation for continuous outcome. Outcome estimation errors were normalized by the range of the outcome. OSRCT allows us to evaluate causal modeling methods on a wide range of data sets for which unbiased effect estimates are available.

data, creating D_{OSAPO} and D_{OSRCT}. For D_{OSRCT}, as described in Theorem 3, outcome estimation was evaluated by weighting the errors in the complementary sample. However, in D_{OSAPO}, no complementary sample is created, since the selected treatment is guaranteed to be observed for every unit. Instead, we can divide D_{OSAPO} into training and test sets. If the RCT-style data is created by sub-sampling treatments equally, by Theorem 2, splitting D_{OSAPO} in half leads to a data set approximately the same size as D_{OSRCT}, allowing for comparison with equal training set size. We estimated errors over 100 trials. Results are shown in Figure 4.4.

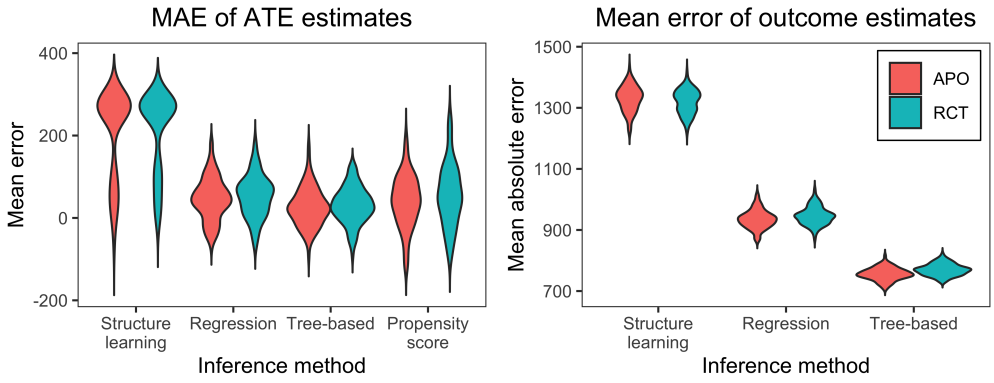

Figure 4.4. APO vs RCT sampling on Postgres data. *Left*: Mean absolute error of ATE estimates, *Right*: Mean error of estimated outcomes The similarity between the RCT and APO data sets suggests that OSRCT and OSAPO produce equivalent constructed observational data.

Results are very similar for the APO data and the RCT-style data constructed from it. Consistent with Theorem 1, this suggests that evaluation with OSRCT data produces equivalent results to OSAPO data. In addition, consistent with Theorem 3, the similarity in outcome estimates suggests that weighting the complementary sample produces equivalent results to an unweighted held-out test set.

4.6 Conclusion

Research progress in machine learning has long depended on high-quality empirical evaluation. Until recently, research in causal modeling has been hindered due to an almost complete lack of 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 modeling methods. They have the benefit of being widely collected by researchers in many fields over many years, and are increasingly being made available for wider use. RCT data is available from a wide variety of domains, and unbiased

estimates of causal effect can be obtained for evaluation. OSRCT can substantially increase the data available for evaluating causal modeling methods.

CHAPTER 5

EVALUATING CURRENT ALGORITHMS FOR CAUSAL MODELING

There are few studies that compare the performance of several causal modeling algorithms, partially due to the lack of standardized data sets for evaluation in causal modeling. Even when data sets are available for comparison, they are either non-interventional or limited in number, limiting the scope of such studies and preventing a nuanced understanding of the relative performance of different approaches. When new algorithms are proposed, they are often only evaluated using structural measures or on data with no ground truth [41, 171, 186, 119], and rarely evaluate on data from more than one source. Such evaluations can be informative about the relative performance of methods on the specific data used, but there is no guarantee that the conclusions drawn will generalize to data from other sources.

To gain a better understanding of the relative performance of causal modeling methods, we perform an evaluation using data sets drawn from the most realistic types and from a variety of sources. As shown in Chapter 4, we can use data from RCTs to evaluate causal modeling methods. RCTs and the computational APO data sets are sources of realistic data that have not been used for large-scale evaluations of causal modeling methods before. We augment these with two additional sources of realistic data: the synthetic-response data sets produced for the ACIC Competition [53] and the IBM Causal Inference Benchmarking Framework [174], and a variety of simulators. This provides us with a range of realistic data, which we can use to perform an evaluation of causal modeling methods. This constitutes a larger-scale evaluation of causal modeling methods than has previously been possible.

We use data from four sources: randomized controlled trials, computational systems, simulators, and data with real covariates and a synthetic response surface. From these categories, we collect 37 data sets and then use them to evaluate seven causal modeling methods. Because the data sets we use generally only have ground truth ATE for a single treatment-outcome pair, we focus this evaluation on methods that estimate bivariate ATE and do not include multivariate structure learning methods.

5.1 Related Work

Although comparative studies are rare, a small number of papers have examined the relative performance of different causal modeling methods. Hahn et al. [83] compare multiple tree and forest based causal modeling methods, using both synthetic data and data from the ACIC competition [82]. They find that modified versions of BART performed better than causal forests and the default BART implementation. In their paper proposing the Dragonnet architecture, Shi et al. [173] compare Dragonnet to five other neural-network-based methods, using data from the ACIC competition and additional synthetic-response data from the Infant Health and Development Program [88]. They find that, on average, Dragonnet net performed the best. Dorie et al. [53] perform a large-scale evaluation using synthetic-response data, comparing algorithms submitted by teams as part of a competition. They find that methods that focused on modeling outcome, even without modeling treatment, generally outperformed methods focused on modeling treatment. They also find that BART (which models outcome) and SuperLearners (ensembles of methods) performed best overall. Elze et al. [57] compare multiple propensity score-based methods: propensity score matching, propensity score stratification, inverse probability of treatment weighting, and including the propensity score as a regression covariate. They use four empirical data sets, for which ground truth was estimated from the original published study. They find that propensity score matching produces the best covariate balance,

propensity score stratification performs well if there is not much covariate imbalance and if the number of strata is chosen carefully, IPTW performs poorly when there were extreme propensity score values (when there is very poor covariate balance), which can be mitigated by augmenting it with a doubly robust method, and that covariate adjust performs well overall.

5.2 Data for Evaluation

We collect data from four general methods of generating data, to create a diverse collection of data sets for evaluation. To be usable for evaluation, it must be possible to calculate an unbiased estimate of treatment effect. For this work, we choose data sets that are initially unbiased (so treatment effect can be estimated), and we induce confounding bias by sub-sampling, as described in Algorithms 1 and 2.

5.2.1 Computational Systems

As described in Chapter 3, Garant and Jensen [64] created three data sets where all potential outcomes are observed. 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, as described in Table 5.1.

5.2.2 Randomized Controlled Trials

We selected data sets from six repositories of RCTs: Dryad [3]; the Yale Institution for Social and Policy Studies Repository [12]; the NIH National Institute on Drug Abuse Data Share Website [5]; the University of Michigan's ICPSR repository [10]; the UK Data Service [9]; and the Knowledge Network for Biocomplexity [7]. We selected these repositories because they contained RCT data, were reasonably well-documented, and had a simple access process. None of these repositories house RCT

data exclusively, so some search and filtering was necessary to identify relevant data sets.

Many other repositories of RCT data exist, but they have higher access restrictions. Access to these other repositories generally involves requesting permission for any desired data set. For some, this request only involves submitting a brief description of the intended use and proving sufficient credentials. For others, this request may require a detailed data analysis plan and description of the benefits of the research. Examples include the National Institute of Diabetes and Digestive and Kidney Diseases [4], Vivli [11], The National Institute of Mental Health Data Archive [8], Project Data Sphere [6], and the Data Observation Network for Earth [2].

The data sets selected for the evaluation met five criteria:

- **Random assignment:** Treatment must be fully randomized for OSRCT to work as intended. We ensured that the selected data sets were created by randomly assigning treatment to each unit.

- **Independent units:** Many causal modeling methods assume independent data instances, so we ensured that the units in the data sets could reasonably be assumed independent (e.g., no spatial correlation).

- **Measured pre-treatment covariate:** At least one measured pre-treatment covariate is necessary to induce confounding bias. The data sets we selected all had multiple pre-treatment covariates, allowing us to select one that was correlated with outcome to induce confounding bias.

- **Reasonably large sample size:** Many RCT data sets are very small ($N < 100$). We selected only reasonably large data sets ($N > 500$).

- **Ease of use:** Some data sets were poorly documented or stored the data over many files. We selected data sets that would require minimal pre-processing.

In cases where treatment was not binary, a reasonable binary version of treatment was constructed, either by grouping merging treatment categories or by selecting a subset of the data with only two values of treatment. Details about these data sets are given in Table 5.1.

5.2.3 Synthetic-Response Data

Many data sets for evaluation were created for the ACIC Competition [53] and the IBM Causal Inference Benchmarking Framework [174]. These data sets were created using a set of real-world covariates and then simulating both a treatment and an outcome. While these are similar to the APO and RCT data described above, in that treatment is generated synthetically based on values of one or more covariates, in these data sets, the outcome 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. Tables 5.2 and 5.3 provide details on the selected data sets.

5.2.4 Simulators

To increase the diversity of data set types, we also generated data sets from three simulators of varying complexity: (1) A simulator of neuropathic pain [192]; (2) Nemo [79], a simulator of population dynamics; and (3) three simple simulators from the WhyNot Python package [139]. 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.

For both the neuropathic pain simulator and the WhyNot simulators, the selected treatment was known from design to be set randomly, and a pre-treatment biasing covariate was chosen based on domain knowledge, producing data sets of the same

Source	ID	Coding	Sample Size	Num Covars	Treatment	Outcome	Biasing Covar
Dryad	4f4qrfj95	RCT-1	6453	27	Temperature	Plant health	Species
Dryad	B8KG77	RCT-2	15289	4	Video type	Bicycle rating	Bike access
HDV	WT4I9N	RCT-3	551	5	Fact truth	Fact removed	Fact cited
ICPSR	20160213	RCT-4	10	5573	Guest race	Accepted	Prior black tenants
ICPSR	23980	RCT-5	10098	7	Age	Resume response	Volunteer service
ISPS	d037	RCT-6	4859	2	Race	Legislator response	Party
ISPS	d084	RCT-7	48509	6	E-mail source	Voter turnout	Prior election turnout
ISPS	d113	RCT-8	10200	4	Mailing	Voter turnout	Gender
KNB	1596312	RCT-9	760	4	Soil heating	C02 levels	Depth
KNB	f1qf8r51t	RCT-10	8063	4	Plant protection	Plant survival	Location
musiclab	-	RCT-11	3719	13	peer-influence	average rating	music knowledge
NIDA	P1S1	RCT-12	776	5	Nicotine levels	Cigarettes per day	Weight
UK Data Service	852874	RCT-13	343	5	Shown video	Response	Ethnicity
UK Data Service	853369	RCT-14	4210	3	Biasing instruction	Line-up identification	Recruitment method
UK Data Service	854092	RCT-15	691	5	Fact check validity	Reaction	Political activity
JDK	-	APO-1	473	5	Obfuscate	Num bytecode ops	Test javadocs
Networking	-	APO-2	2599	1	Proxy	Elapsed time	Server class
Postgres	-	APO-3	11128	8	Index level	Runtime	Rows returned
Nemo	-	Sim-1	10000	9	Breeding	Adult viability	Deleterious loci
Nemo	-	Sim-2	10000	9	Deleterious model	Deleterious frequency	Mutation rate
Nemo	-	Sim-3	10000	10	Dispersal rate	Survival	Deleterious loci
Neuropathic pain	-	Sim-4	10000	25	DLS L4-L5	Lumbago	DLS L5-S1
Neuropathic pain	-	Sim-5	10000	25	DLS C5-C6	Right Skull pain	DLS C3-C4
Neuropathic pain	-	Sim-6	10000	25	DLS C4-C5	Right Shoulder pain	DLS C6-C7
WhyNot	opiod	Sim-7	10000	3	Abuse	Overdose deaths	Illicit users
WhyNot	world2	Sim-8	10000	6	Capital investiment	Population	Pollution
WhyNot	zika	Sim-9	10000	9	Zika control strategy	Symptomatic humans	Exposed mosquitoes

Table 5.1. Data sets used in experiments. 'ID' denotes the repository-specific ID for each data set, where applicable. 'Coding' denotes the shortened data set name used in figures.

format as RCTs. For Nemo, it was possible to run parallel simulations with different treatment values, producing APO data sets. Table 5.1 provides details about the simulator data sets.

5.3 Algorithms to Compare

Due to the nature of the ground truth in most of our data sets (treatment effect of a single treatment on a single outcome), we focused our evaluation on causal modeling methods that estimate average treatment effect. We chose seven methods to evaluate: propensity score matching (PSM), inverse probability of treatment weighting (IPTW) [158], outcome regression (OR), Bayesian additive regression trees (BART) [43], causal forests (CF) [193], doubly-robust estimation (DRE) [60], and

ID	Coding	Sample Size	Num Covars	Treatment Function	Percent Treated	Outcome Function	Alignment	Treatment Effect Heterogeneity
4	SR-1	4802	56	Polynomial	35%	Exponential	75%	high
27	SR-2	4802	56	Polynomial	35%	Step	25%	Medium
47	SR-3	4802	56	Polynomial	65%	Exponential	75%	High
65	SR-4	4802	56	Step	65%	Step	75%	Medium
71	SR-5	4802	56	Step	65%	Step	25%	High

Table 5.2. ACIC Data sets used in experiments. 'ID' denotes the ACIC ID for each data set. 'Coding' denotes the shortened data set name used in figures.

ID	Coding	Sample Size	Num Covars	Percent Treated	Effect Size	Link Type
1b50aae9f0e34b03bdf03ac195a5e7e9	SR-6	10000	151	69%	-3.2	Polynomial
2b6d1d419de94f049d98c755beea4ae2	SR-7	10000	151	23%	-0.13	Log
19e667b985624159bae940919078d55f	SR-8	10000	151	17%	0.06	Exponential
7510d73712fe40588acdb129ea58339b	SR-9	10000	151	27%	0.017	Log
c55cbee849534815ba80980975c4340b	SR-10	10000	151	19%	-0.23	Exponential

Table 5.3. IBM Data sets used in experiments. 'ID' denotes the IBM ID for each data set. 'Coding' denotes the shortened data set name used in figures.

a neural-network-based method, Dragonnet (NN) [173]. Chapter 1 provides details about these methods.

5.4 Experimental Setup

For each data set, we calculated the unbiased ATE to use as a ground truth. We then sub-sampled, as shown in Algorithms 1 and 2, to produced a biased data set. All data points rejected by the sampling were held out as the complementary sample, which was used for evaluating outcome prediction, as described in Section 4.2.3. All algorithms were applied to the biased data, producing estimates of ATE. All algorithms that are capable of predicting individual-level outcomes also produced predicted outcomes for the complementary sample. 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. The only exception was the five

data sets from the IBM Causal Inference Benchmarking Framework, for which we sampled to 5,000, due to their high dimensionality.

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.

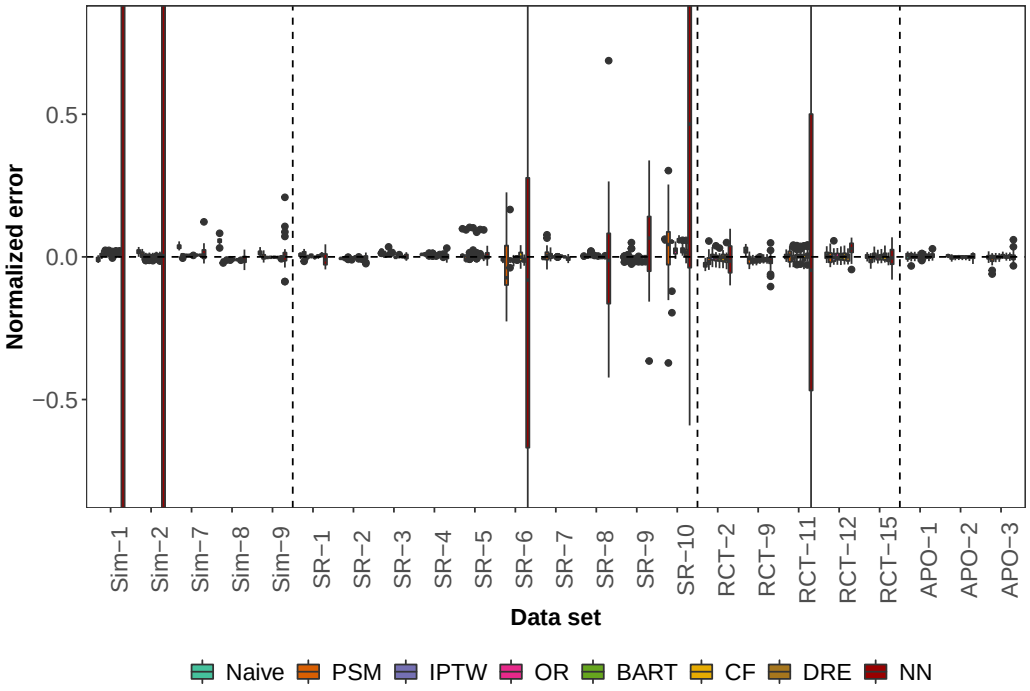

Figure 5.1. Normalized error in estimating ATE: variability for the neural network method is extremely high

5.5 Results

Effect estimation 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

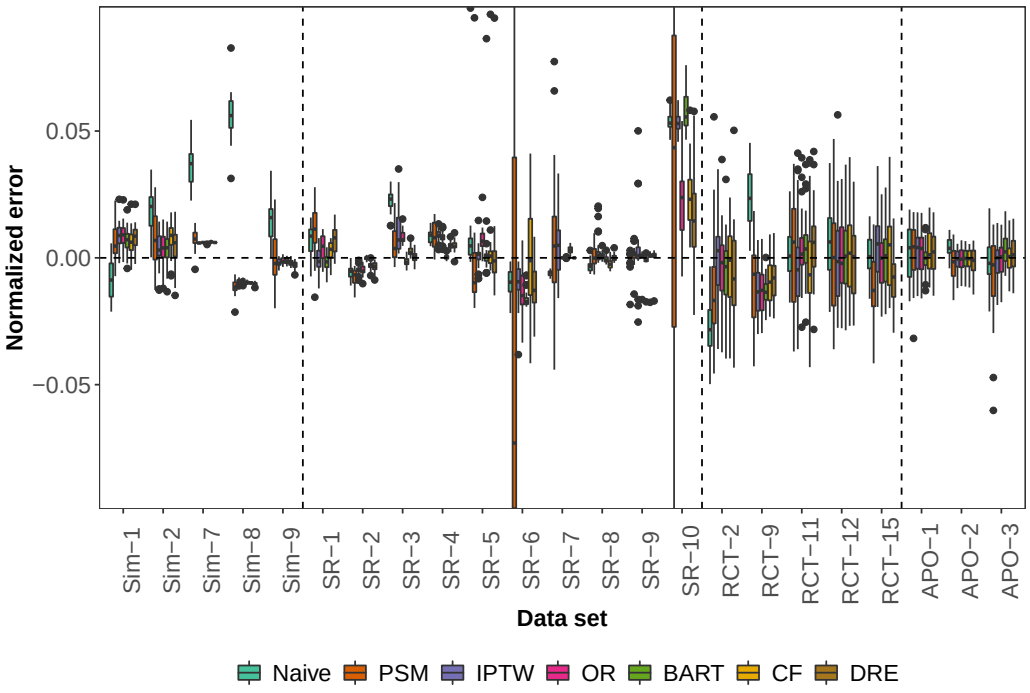

Figure 5.2. Normalized error in estimating ATE, without the neural network method: variability for propensity score matching is higher than for other methods

data sources, though, can also provide a useful understanding of how different data design choices affect estimates of relative performance. Initial results for the experimental setup described above, for data sets with continuous outcomes, are shown in Figure 5.1.

The primary notable feature of these results is the extremely high variability of the neural network method. There are at least two possible reasons for this. As we initialize different random weights in each run, the model might be sensitive to the initialization weights and converge to different local optima. In addition, sample size for most of the data sets is less than 5000, which is significantly lower than is typically used for neural network based methods. This might produce overfitting and thus high variability. For ease of visualization, this method was omitted from the rest of the graphics in this section.

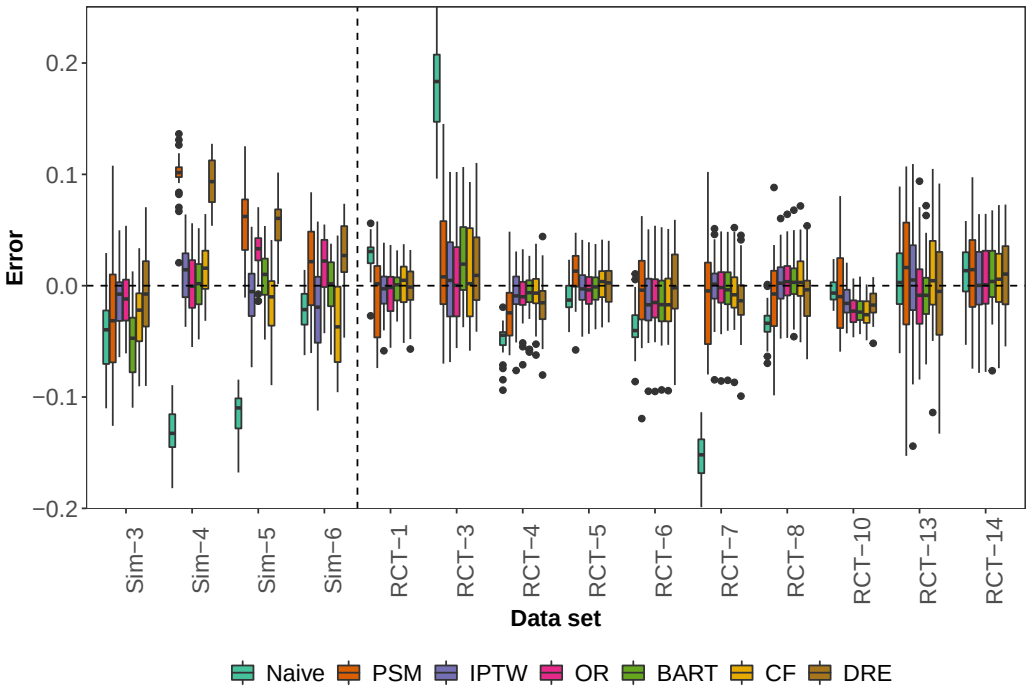

Figure 5.3. Normalized error in estimating risk difference: most methods have very similar performance, though propensity score matching often has higher variability.

Figures 5.2 and 5.3 show results without the neural network method, for data sets with continuous outcome and binary outcome, respectively. Without the neural network method, results across data sets appear fairly similar. The only exception is propensity score matching, which 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, by matching solely on the propensity score [108]. For ease of comparison of the remaining methods, we omit propensity score matching from the remaining ATE results, which are shown in Figure 5.4.

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

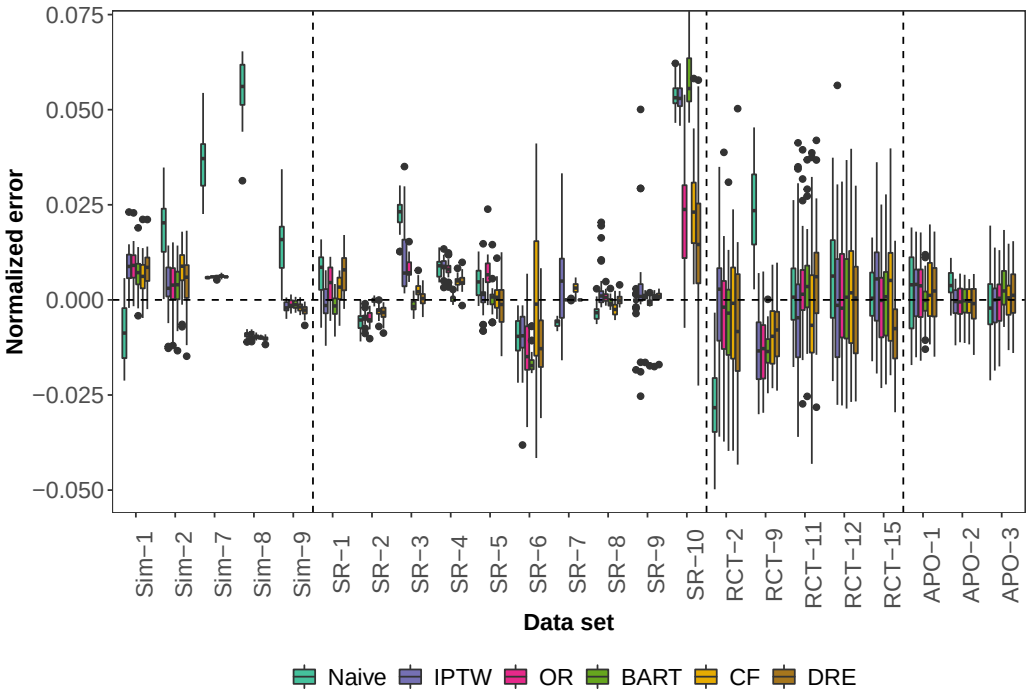

Figure 5.4. Normalized error in estimating ATE, without propensity-score matching: for the RCT, APO, and simulator data sets, most methods have very similar performance. However, performance varies more for the synthetic-response data sets

in contrast to the synthetic-response data sets, where we see far more variability between methods on the same data set. This contrast between the synthetic-response data sets and the other three types has many 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 not realistic.

Another possible explanation, however, is that, while the response surface in the RCT data sets is realistic, the treatment assignment is very simplistic, 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

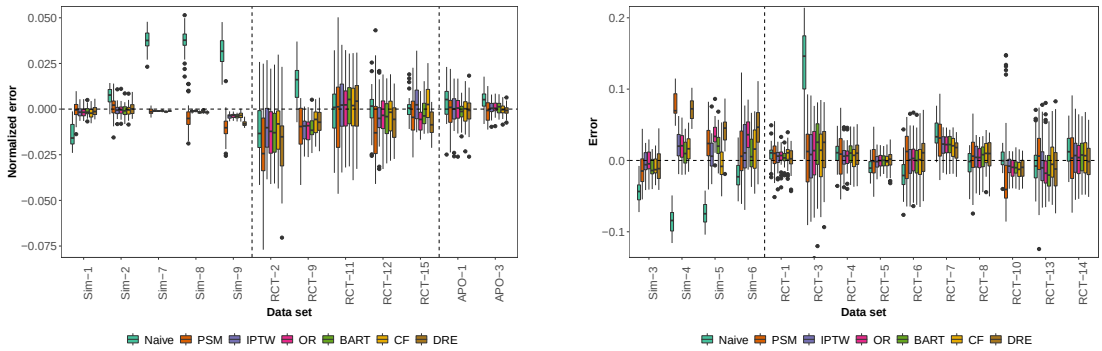

Figure 5.5. Left: Normalized error in estimating ATE, Right: Error in estimating risk difference, with two biasing covariates: performance is very similar as with a single biasing covariate

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 were used.

Results with two biasing covariates are shown in Figure 5.5. 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, it is also possible that the complexity of the response surface in the synthetic-response data sets is more complicated than is necessary to approximate many instances of real-world causal effects.

Another potential hypothesis for the performance difference for synthetic-response data sets is dimensionality. The synthetic-response data sets have significantly more variables than the other data sets, and this high dimensionality may be leading to more varied performance between methods. To test this, we reduced the number of variables for the ACIC competition data sets, from the original 56 down to 17. The 17 variables were chosen to be a super-set of the variables included in the treatment

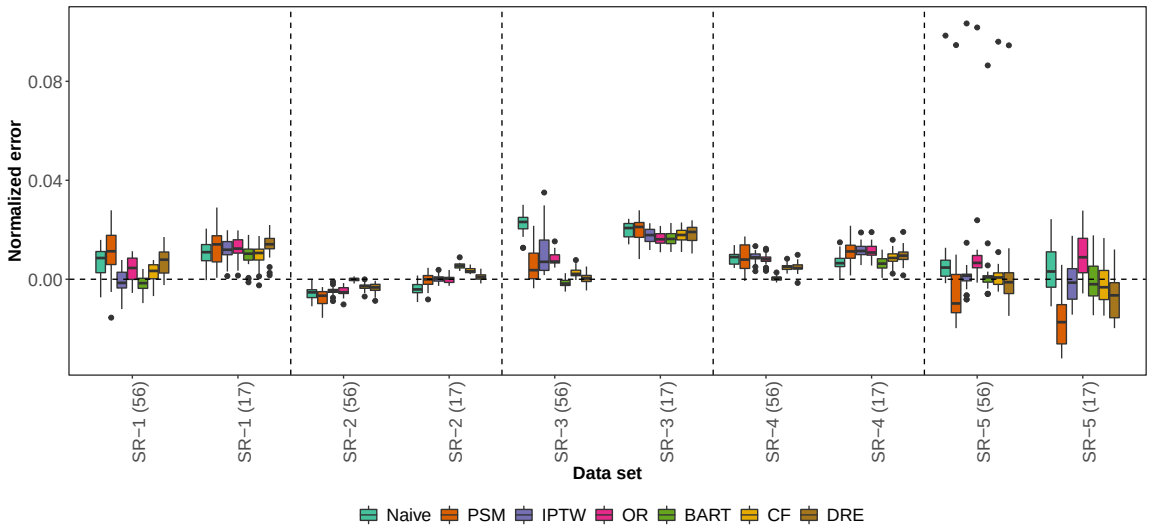

Figure 5.6. Dimensionality comparison for the ACIC competition data sets. With SR-1 and SR-6, performance between methods is more similar with 17 variables than with 56

and response functions. We then compared the performance of the causal modeling methods between these two groups of data sets. Results are shown in Figure 5.6 and suggest that this hypothesis may be partially correct. For at least two of the five ACIC competition data sets, when the number of variables is reduced to 17, performance between different algorithms becomes a lot more similar, appearing more in line with the results for the RCT data sets.

Another aspect of performance that we can investigate with this data is the effect of sample size on performance. To investigate this, we ran some additional experiments, varying the sample size from 1000 to 10000. We restricted the focus of these experiments to the simulators, since we could generate any number of samples from them. Results of these experiments for BART are shown in Figure 5.7. Overall, these results match our expectation: as sample size increases, variability decreases, and the mean error approaches zero. These results for BART are similar to those for all other algorithms, with the exception of propensity-score matching, where performance is consistently poor across all sample sizes.

Figure 5.7. Left: Normalized error in estimating ATE, Right: Error in estimating risk difference, as sample sizes increases for BART: as sample size increases, variability decreases and error approaches zero

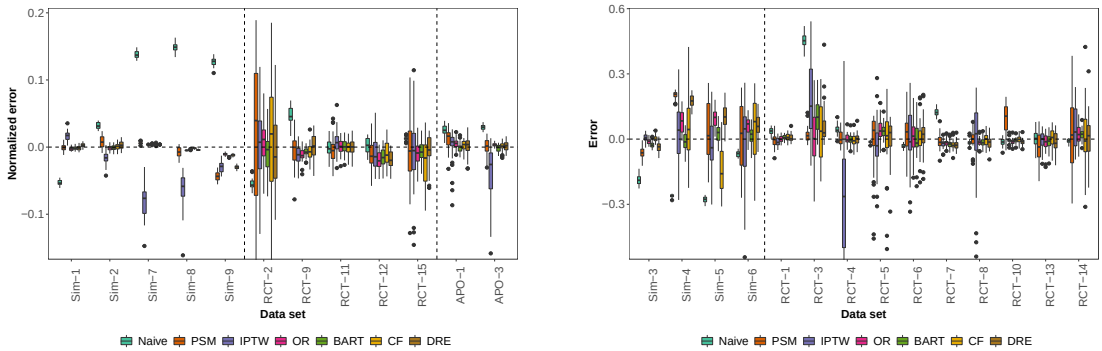

Figure 5.8. Left: Normalized error in estimating ATE, Right: Error in estimating risk difference, with two biasing covariates and increased bias strength: with higher bias strength, IPTW does significantly worse

We also tested algorithm performance for different levels of bias. Figure 5.8 shows results with two biasing covariates, with significantly higher biasing strength, increasing the degree to which the biasing covariates determine treatment. With this degree of bias, for the simulator data sets, IPTW and causal forest, which had previously performed similarly to other methods, now are significantly worse. The likely reason for this poor performance is that, when bias gets very high, overlap is lost, with the values of the biasing covariates for the treatment group almost totally distinct from the values in the control group. This lack of overlap is a violation of the positivity assumption, an assumption made by many causal modeling algorithms. IPTW and causal forest both focus on estimating the probability of treatment, so it makes sense that these methods would suffer the most in this situation.

While the ranges of variability for most methods are the same, this doesn't guarantee that each method is producing the same result for each of the 30 trials. Each methods error could be uncorrelated with the others, suggesting that an ensemble approach might improve performance. To test this, we computed a correlation matrix for each data set, calculating correlation across the 30 trials for each method. Results for a few representative data sets are shown in Figure 5.9. In most cases, the correlation is the weakest with the neural network method, and is generally weaker with propensity score matching. For all other methods, though, errors are highly correlated. There are some exceptions, as in SR-7. The reason for these varies. In the case of SR-7, this is likely a result of the low variability across the 30 trials. Correlation matrices for all other data sets can be found in Appendix B.2.

Figure 5.10 shows overall mean performance for each algorithm. As observed above, propensity score matching has the highest error overall. In addition, doubly-robust estimation appears to have higher error for data sets with binary outcomes. More nuance can be seen in Figure 5.11, which shows mean error by data source. The higher error for doubly robust estimation appears to be primarily for simulator

Figure 5.9. Correlation matrices for four data sets. In most cases, error is highly correlated

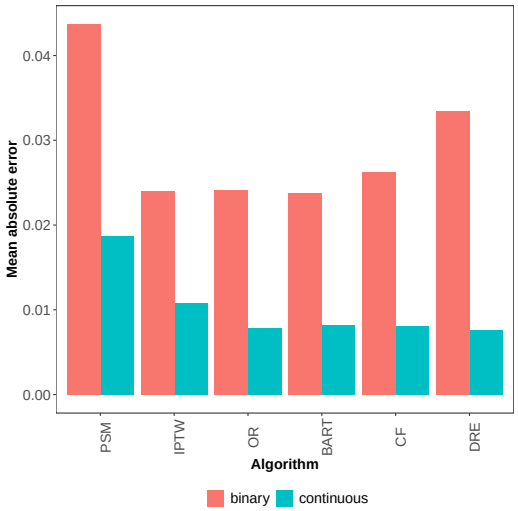

Figure 5.10. Overall mean absolute error by algorithm

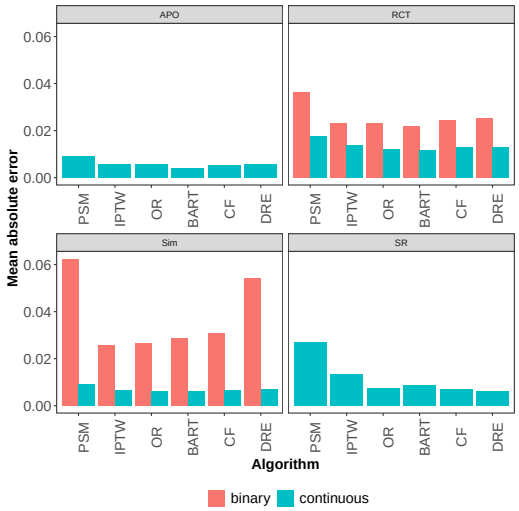

Figure 5.11. Overall mean absolute error by algorithm, by source of data

data sets. For the other data sources, mean performance is fairly consistent across algorithms.

As discussed in Section 4.2.3, when using OSRCT, the complementary sample can be used to evaluate algorithms on their ability to estimate individual-level outcomes. For APO data sets and synthetic-response data sets, a held out test set can be used instead, which, by Theorem 3, is equivalent to using the weighted complementary sample. However, many of the algorithms we are evaluating here are not capable of producing individual-level outcome estimates, so this evaluation is limited to only BART and outcome regression. Results for outcome estimation are shown in Figures 5.12 and 5.13.

Unsurprisingly, BART consistently outperforms outcome regression. Both of these methods focus on modeling the response surface, but BART uses a higher capacity tree-based model rather than a simple regression. The difference is far stronger for data sets with a continuous outcome, compared to those with a binary outcome. The difference is also minimal for the RCT data sets. This trend is constant as we increase the strength of the biasing and when two biasing covariates are used.

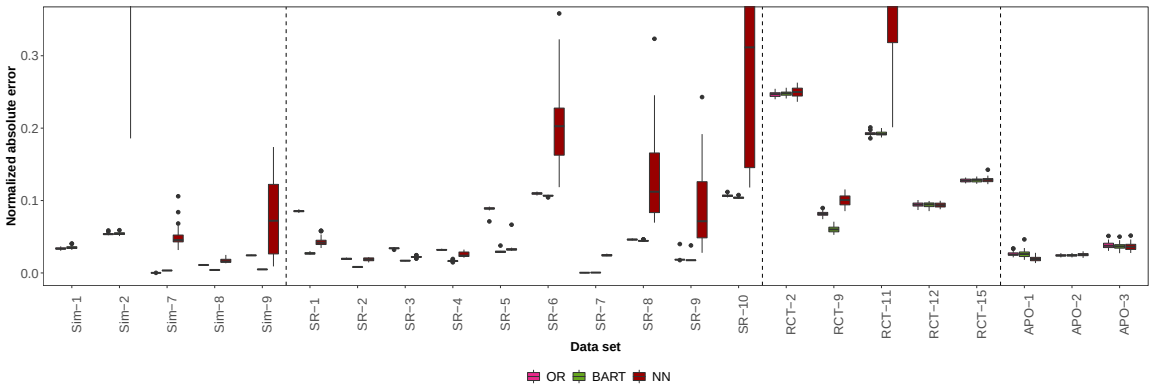

Figure 5.12. Normalized absolute error in estimating a continuous outcome: BART generally outperforms outcome regression, and the neural network method sometimes has extremely high variability

The neural network method performs worse overall, often with high variability. This evaluation is unfortunately limited since none of the other algorithms we evaluated are capable of producing individual-level outcome estimates. In general, methods that model outcome are more likely to provide this as an option, making this a useful evaluation tool when comparing multiple outcome estimation-based methods.

In summary, this analysis supports several conclusions. Performance with the neural network method has extremely high variability. This could be a result of inherent randomness in the neural network method, leading to more variability between runs than the other methods. Alternatively, this could also be due to a lack of hyper-parameter tuning or the small data set sizes. Performance with propensity score matching also has higher variability. For all other methods, performance is similar when using the RCT, APO, and simulator data sets, while performance varies more for the synthetic-response data sets. This variability could be a result of the more complicated outcome function, the more complicated treatment function, or high dimensionality. We saw no change in results when increasing the complexity of the treatment function for other data sets, and methods that focus on estimating treatment aren't consistently performing better, so the complexity of the treatment

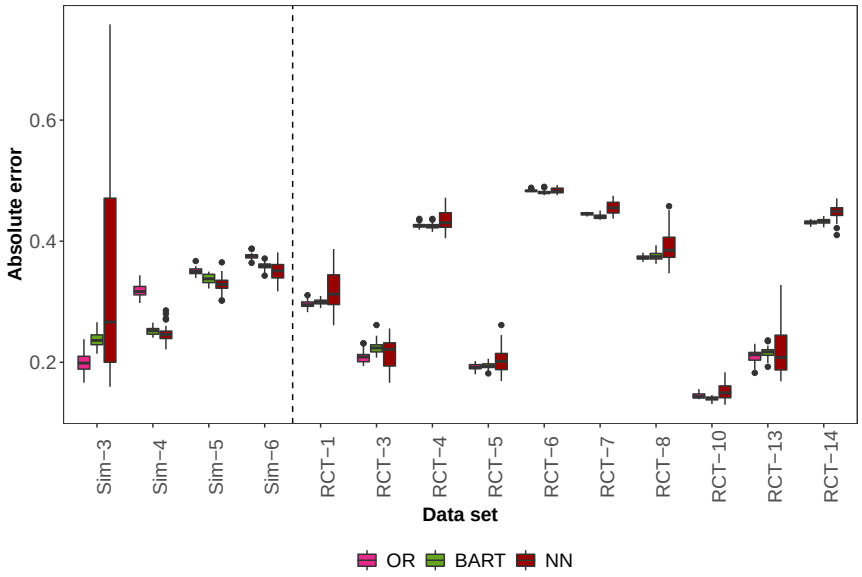

Figure 5.13. Error in estimating a binary outcome: BART and outcome regression have similar performance, and the neural network method sometimes has extremely high variability

function may not be the answer. However, we showed some preliminary results suggesting that the dimensionality may play a role.

5.6 Future work

There are many interesting directions to extend this work. One of the key takeaways from this analysis is that, overall, most methods have very similar performance. However, by looking at a wider variety of data sets, it may be possible to find situations where some data sets out-perform others. For example, a larger number of RCTs could be collected with different features (binary or continuous outcomes, many or few covariates, discrete or continuous features, strong or weak treatment effect), which could be used to test if some methods perform better than others in different specific situations.

Similarly, more features of the experimental setup could be varied to test performance in a variety of situations. We tried varying the sample size just for the

simulators, but sample size could also be varied for the synthetic-response data sets and for some of the larger RCTs. Some methods may be more efficient with smaller sample sizes, or may have better performance with very large sample sizes.

There are also opportunities to further investigate the differences between the synthetic-response data sets and the others. We performed some experiments increasing the complexity of the treatment function, from a single biasing covariate to two biasing covariates, but there was no significant difference. However, it's possible that increasing the number of confounders even more would make the RCT data sets produce similar results to the synthetic-response data sets. This analysis was partially limited by the number of covariates in the RCTs, some of which were very limited, as well as the diversity of data types in the RCT covariates, preventing us from applying a single biasing function across many data sets. A larger number of RCTs, with higher dimensionality, could be collected, and more complicated treatment biasing functions defined, to help test if bias function complexity is a contributing factor to the performance variability observed for the synthetic-response data sets. Such a set of RCTs could also be used to further test the hypothesis that the performance difference is due to the larger number of covariates in the synthetic-response data sets.

Another avenue for investigation is fine-tuning specific methods. This could be especially beneficial for the neural network method, where the extremely high variability in performance could be the result of poor parameter settings. Other methods could also benefit from some data set-specific tuning. In this analysis, each method was implemented using its default settings, but it may be possible to increase performance by varying these settings. Future work could look at not just the default performance of these methods, but the best case performance.

Finally, more analysis could be done on evaluating outcome estimation using the complementary sample. This type of analysis was limited due to our choices of meth-

ods to evaluate, since many were not capable of producing individual-level outcome estimates. Many outcome estimation-based methods for causal modeling do exist, though, and a deeper evaluation of those types of methods, using both ATE estimation and outcome estimation, could be informative. This could also allow for investigation of how important individual-level outcome estimates are to overall treatment effect estimation, by assessing how well outcome estimation performance predicts ATE estimation performance.

5.7 Conclusions

We have performed an evaluation of seven causal modeling methods over 37 data sets. These data sets are drawn from four sources of empirical data for causal model evaluation, two of which have never been used before for a large-scale evaluation. Our results suggest the importance of data diversity for evaluation. We found that data from RCTs and the three APO data sets from computational systems produce similar results. However, this performance differs somewhat from simulators, and significantly from synthetic-response data. The varying performance by data source demonstrates the importance of evaluating on a variety of data. Many evaluations of causal modeling methods only evaluate on synthetic data and one or two empirical data sets. This type of evaluation shows a method's performance for only the specific properties of the chosen data set, leaving our understanding of the method's performance obscured.

This work opens up multiple possible paths for future exploration. The difference in performance between RCT and synthetic-response data sets should be evaluated in greater detail, by increasing the complexity of the biasing functions for the RCTs to match the complexity of the synthetic-response data sets. More work could be done to test additional causal modeling methods and different implementations. For ease of comparison, we only used the default implementations of the algorithms, and

more fine-tuned implementations (including parameter tuning for the neural network method) could lead to some interesting performance differences. Finally, other types of data could be compared against. Our analysis included data from simulators, but many papers evaluate on purely synthetic data (such as from a hand-specified structural equation model), and it would be interesting to see how results on synthetic data compare to more empirical data. We hope that this evaluation can motivate others in the field to undertake similar projects with different focuses, giving us a clearer picture of the relative performance of causal modeling methods than has previously been possible.

CHAPTER 6

CONCLUSIONS

Evaluation is an important component of any field of science. It affects how we perceive methods and what research directions we choose to pursue, and it affects how we make decisions about deploying methods in practice. Until recently, evaluation within the causal modeling community has been severely limited due to a dearth of empirical data resources. There is an increasing effort to produce data sets that can be used for evaluation. However, these data sets require a great deal of work to collect, so while the data available for evaluation is increasing, it still lags significantly behind other fields of machine learning.

In Chapter 3, we demonstrated the value empirical data and interventional measures can provide for evaluating causal modeling algorithms. Empirical data lets us assess how well a method works in realistic situations, which is important if we expect anyone to use these methods in practice. Interventional measures let us assess how well a method can estimate the effects of actual interventions, which is important if we expect anyone to use these methods for making decisions.

We are not the first to point out the need for more robust evaluation techniques. Some of the data sets we discuss were created in response to recognition that better evaluation was necessary [53, 174, 143]. In addition, prior work has examined the importance of testing the generalizability of causal modelings drawn from observational data [218, 106] and comparing causal effects drawn from observational and experimental data [46, 56, 55, 76]. However, despite this, as our survey shows,

empirical evaluation with interventional measures is rarely used by computer science researchers.

To aid in increasing the availability of data for evaluation, in Chapter 4, we described a method for using data from randomized controlled trials to evaluate causal modeling methods. RCTs have the benefit of being widely collected by researchers in many fields over many years, and are increasingly being made available for wider use. RCT data is available from a wide variety of domains, and unbiased estimates of causal effect can be obtained for evaluation.

In Chapter 5, using a combination of empirical data and interventional measures, we performed a large-scale evaluation using a larger number of data sets than was previously possible. Our results emphasize the importance of evaluating on data from multiple sources. The difference in algorithm performance between the synthetic-response data sets and the others suggests one of two things: that high levels of complexity is not necessary to create realistic response surfaces, or that more complicated treatment biasing functions can provide a better understanding of the relative performance of methods. Future work should look to investigate this. We also found that propensity score matching, as the literature suggests, often performs poorly in practice, and that care should be taken to control for variability if applying a neural network based method for causal modeling.

Overall, current standard practice in the evaluation of causal modeling methods is largely insufficient, and improving our evaluation methods can have a strong positive effect on the community. The increasing availability of empirical data for causal model evaluation means that minimalist empirical evaluation should no longer be acceptable. By holding our evaluations to a higher standard, we not only improve our own understanding of causal modeling methods, guiding future research direction, but we also demonstrate the effectiveness of our methods to those outside the community, increasing the adoption of methods that can be shown to perform well.

APPENDIX A

ACRONYMS

Acronym	Meaning
ACIC	Atlantic Causal Inference Conference
APO	All Potential Outcomes
ATE	Average Treatment Effect
BART	Bayesian Additive Regression Trees
CF	Causal Forests
CGM	Causal Graphical Model
DAG	Directed Acyclic Graph
DRE	Doubly-Robust Estimation
GES	Greedy Equivalence Search
IPTW	Inverse Probability of Treatment Weighting
MMHC	Min-Max Hill Climbing
NN	Neural Networks
OR	Outcome Regression
OSAPO	Observational Sampling from APO Data
OSRCT	Observational Sampling from RCT Data
PSM	Propensity score matching
RCT	Randomized Controlled Trial
RMSE	Root Mean Square Error
SHD	Structural Hamming Distance
SID	Structural Intervention Distance
TVD	Total Variation Distance

Table A.1. Acronyms used throughout this dissertation

APPENDIX B

ADDITIONAL DETAILS ON COMPUTATIONAL SYSTEM DATA AND ADDITIONAL EXPERIMENTS

B.1 Additional Details on Computational System Data

We introduce a source of empirical data where interventions are possible: large-scale computational systems. We performed experiments on three large computational systems: Postgres, the Java Development Kit, and HTTP processing. These systems have many desirable properties for the purposes of empirical evaluation: (1) They are pre-existing systems created by people other than the researchers for a purpose other than evaluating algorithms for causal discovery; (2) They produce non-deterministic experimental results due to latent variables and natural stochasticity; (3) System parameters provide natural treatment variables; and (4) Each experiment is recoverable, allowing the same experiment to be performed multiple times with different combinations of interventions.

Within each computational system, we measure three classes of variables: outcomes, treatments, and subject covariates. Here, outcomes are measurements of the result of a computational process, treatments correspond to system configurations and are selected such that they could plausibly induce changes in outcomes, and subject covariates logically exist prior to treatment and are invariant with respect to treatment. Using these variables, we can apply all combinations of treatments to all subjects, and we can use these results to estimate actual interventional distributions for the effects of each treatment variable on each outcome variable. We can also then sub-sample these experimental data sets in a manner which simulates observational

bias to produce observational-style data sets, allowing us to evaluate an algorithm's performance on pseudo-observational data and evaluate it using actual interventional effects. These data sets will be made available after publication.

We had a number of goals in mind when gathering data from our real domains:

- **Causal Sufficiency:** The algorithms we studied require that no pair of variables in the model are both caused by a latent variable. We can guarantee this is true for pairs of treatments and outcomes (since treatments have no parents in the original data set), but needed to employ domain knowledge to limit sources of causal sufficiency violations with regard to other pairs of variables.

- **Acyclicity:** Each of the systems can be described by a "single-shot" computational process which starts and finishes without the possibility for feedback.

- **Instance Independence:** We took efforts to ensure that each execution of the computational process was independent of previous executions. In most cases, this required clearing caches and resetting other aspects of system state.

- **Plausible Dependence:** We selected variables that we believed would be causally related.

Each domain is characterized by three classes of variables: subject covariates, treatments, and outcomes. Under the factorial experiment design, outcomes were measured for every combination of subjects and treatments. This yields a data set with many records for the same subject, as in the example in Table B.1. To permit greater opportunities for observational sampling, we performed multiple trials of each factorial experiment. Given the difficulty associated with modeling highly complicated outcomes such as runtime, we employed a normalization scheme for each data set, dividing outcome values by a "baseline" value—the median control-case outcome value. Thus, we ultimately recorded outcomes which represent a deviation from this baseline. In this regard, our experimental results resemble a within-subjects design [78],

Subject ID	Covariate	Treatment	Outcome
1	A	0	1.33
1	A	1	0.96
2	B	0	1.89
2	B	1	0.54
3	A	0	1.02
3	A	1	0.99
4	A	0	1.35
4	A	1	1.12

Table B.1. An example of a factorial experiment with four subjects and a binary treatment

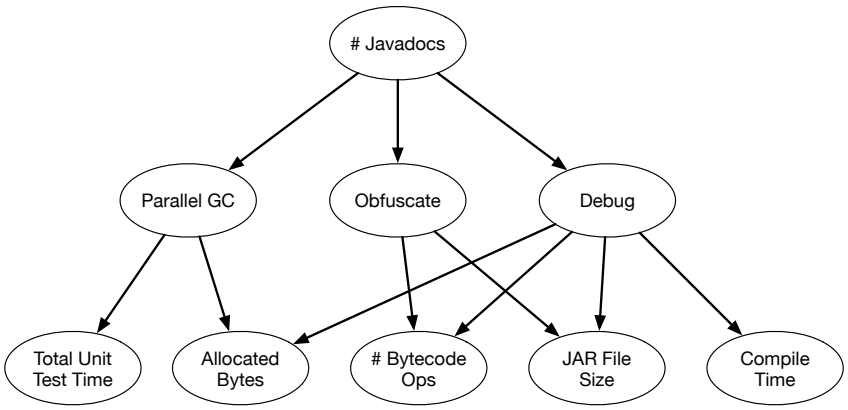

Figure B.1. Consistent model for the JDK domain

although without many of the pitfalls that plague experiments on humans, such as non-independence of outcome measurements. In the original data from each domain, subject covariates are either discrete, continuous, or binary; treatments are binary; and outcomes are continuous. We converted each of the variables to a discrete representation to make parameterization and modeling more robust.

B.1.1 Java Development Kit

Our experiments on the Java Development Kit (version `1.7.0_60`) used 2,500 Java projects obtained from GitHub as the subjects under study. We retrieved only projects which use the Maven build tool to facilitate automated compilation and

execution. Additionally, we constrained our search to include only projects which had unit tests. This may introduce selection bias in our data collection processes, but this is acceptable. It is not important that our conclusions generalize to some population of computational systems, only that there are causal dependencies which hold on the sub-population under investigation. Of those, 473 compiled and ran without intervention. This group yielded a total of 7,568 subject-treatment combinations. For each combination, we compile and execute the unit tests of the Java project. In order to obtain full state recovery between each trial, any compiled project files were cleared between executions. Thirty-five CPU days were required to collect this data using several Amazon EC2 instances.

B.1.1.1 Treatments

- **Aggressive Compiler Optimization**: Disabling this option (enabled by default) prevents some compiler optimizations from running, potentially slowing down execution time but perhaps reducing compilation time. This option is disabled with the `javac` option `-XX:+AggressiveOpts`.

- **Emission of Debugging Symbols**: Debugging symbols are used to provide a map through the compiled source code that can be used for interactive debugging and diagnostics. Inclusion of these symbols may require some time during the compilation phase, increase the size of the compiled program, and could possibly impact runtime. This corresponds to the `-g` flag of `javac`.

- **Garbage Collection Methodology**: The Java Development Kit supports several garbage collection schemes. Two were considered: parallel and serial. These schemes are activated with the `-XX:-UseParallelGC` or `-XX:-UseSerialGC` arguments.

- **Code Obfuscation**: Several third-party tools are capable of obfuscating compiled code, making reverse-engineering difficult. This process could also affect the size of the compiled project files. The yGuard[1] tool was used for this purpose.

B.1.1.2 Outcomes

- **Number of Bytecode Instructions**: Before execution, Java code is compiled to an intermediate language referred to as bytecode. We measured the number of atomic instructions, or operations, in this compiled code to form this outcome using a custom-built bytecode analysis tool based on Javassist[2].

- **Total Unit Test Time**: Each project we gathered contains one or more unit tests. To capture the runtime of the full unit test workload, we computed the sum of runtimes of all unit tests for a given project.

- **Allocated Bytes**: The Java Virtual Machine supports a profiling option (`-agentlib:hprof=heap=sites`) which can be used to track heap statistics throughout a program's execution. We utilized this feature to obtain the total number of bytes allocated during unit test execution.

- **Compiled Code Size**: Java programs are often packaged in an format known as a JAR (Java ARchive). To characterize the size of the compiled code, we recorded the size in bytes of the associated JAR file.

- **Compilation Time**: In order to execute unit tests, the entire project needs to be compiled. This outcome represents the time used to convert all source files to their bytecode equivalents.

---

[1] http://www.yworks.com/en/products_yguard_about.html

[2] http://www.csg.ci.i.u-tokyo.ac.jp/ chiba/javassist/

B.1.1.3 Subject Covariates

All subject covariates were obtained using the JavaNCSS tool[3].

- **# NCSS (non-comment source statements) in Project Source**: This covariate is highly predictive of compiled code size. Conceivably, in observational settings, large projects could also be associated with more liberal use of advanced compilation settings and tools, such as a code obfuscator.

- **# NCSS, Functions, and Classes in Unit Test Source**: These covariates are somewhat representative of the unit test workload. Projects with many lengthy unit tests may also have longer total unit test runtime.

- **# "Javadoc" comments in Unit Test Source**: This covariate could be indicative of code quality. Well-commented code is perhaps more likely to be found in high-quality projects. This code may be more likely to be used in production environments, and thus could be less likely to be observed with debugging symbols. This feature is used in the treatment-biasing procedure for construction of observational data sets.

B.1.2 Postgres

Consistent with a data warehousing scenario, we employ a fixed database for our Postgres (version 9.2.2) experiments: a sample of the data from Stack Overflow, drawn from the Stack Exchange Data Explorer[4]. The data explorer also houses many user-generated queries. We collected 29,375 of the most popular queries to use as subjects for this study. Stack Exchange's data warehouse uses Microsoft SQL Server, which does not completely overlap with Postgres in supported features and syntax. Some queries use only ANSI-compliant syntax and run successfully on either SQL Server or

[3]http://javancss.codehaus.org/

[4]http://data.stackexchange.com/

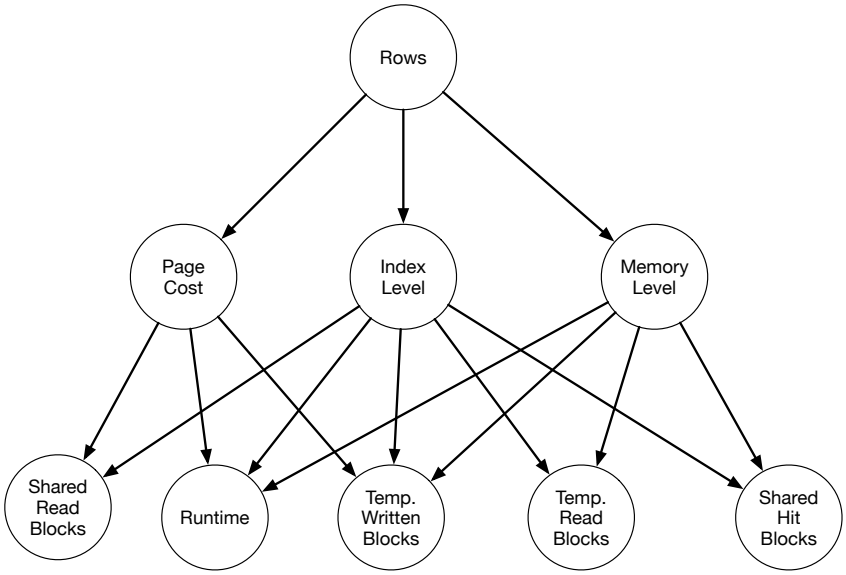

Figure B.2. Consistent model for the postgres domain

Postgres. To obtain as large a set of subjects as possible, we employed a semantics-preserving query rewriting scheme to adapt queries into Postgres-compliant syntax wherever possible. This yielded a set of 11,252 user-generated queries which executed successfully within Postgres for a total of 90,016 subject-treatment combinations. In order to recover system state between trials, the shared memory setting (specifying how much main memory Postgres can use for caching) was set to 128 kilobytes, limiting caching significantly. Any queries which required more than 30 seconds to execute were marked as "failures" in order to prevent long-running queries from holding up other queries, which typically required one second to execute. As with the JDK data set, this may induce sampling bias, but we are not aiming for our experimental findings to generalize to the broader population of database queries.

B.1.2.1 Treatments

- **Indexing**: A common administration task is to identify indices that can be used to accelerate lookup of commonly-referenced columns with a particular value or falling within a range. For our experiments, we employed two indexing

settings: no indexing, and indexing on primary key/foreign key fields. Domain knowledge suggests that that the latter approach would dramatically reduce runtime of some queries. In all cases, the default B-tree index was employed.

- **Page Cost Estimates**: In order to determine if an index should be used, the database employs estimates of the relative cost of sequentially accessing disk pages and randomly accessing disk pages. We utilized two extremes for this setting: one scheme in which random page access is estimated to be fast, relative to the sequential page access, and one scheme in which the opposite relation holds. The corresponding database settings we adjusted were `random_page_cost` and `seq_page_cost`.

- **Working Memory Allocation**: The database engine can make use of fast random-access memory, if available, to store intermediate query results. The amount of working memory that is allocated to the system can be controlled with a configuration option. For our investigation, we employed a low-memory setting and a high-memory setting, with background knowledge suggesting that the latter would result in faster-executing queries. This treatment was instrumented with the `work_mem` and `temp_buffers` options.

B.1.2.2 Outcomes

- **Blocks Read from Shared and Temporary Memory**: These two outcomes identify the number of blocks, or memory regions, that were read during query execution. Shared memory is persistent (disk) and is accessed during normal table-retrieval procedures. Temporary memory is volatile (main memory) and is used for staging ordering or joining operations.

- **Blocks Hit in Shared Memory Cache**: This outcome represents the number of memory reads that were to be performed against shared memory, but were identified instead in a main memory cache.

- **Runtime**: The total time to execute the query.

B.1.2.3 Subject Covariates

- **Year of Query Creation**: The year that the query was entered on the Stack Exchange data explorer.

- **Number of Referenced Tables**: The number of distinct tables that are referenced in the query.

- **Total Number of Rows in Referenced Tables**: The sum of cardinalities of tables referenced in the query.

- **Number of Join Operators**: The number of join operators employed in the query, requiring merging data from two tables.

- **Number of Grouping Operators**: The number of grouping operators employed in the query, requiring reduction and possibly summarization of the data.

- **Number of Other Queries Created by the Same User**: The total number of queries that the Stack Exchange user has created.

- **Length of the Query in Characters**: The length of the query after application of relevant rewrite rules.

- **Number of Rows Retrieved**: The number of rows that are returned by the query. Logically, this value exists prior to application of any treatment and is invariant with respect to treatment (since the database is fixed), even though we can only measure it after query execution.

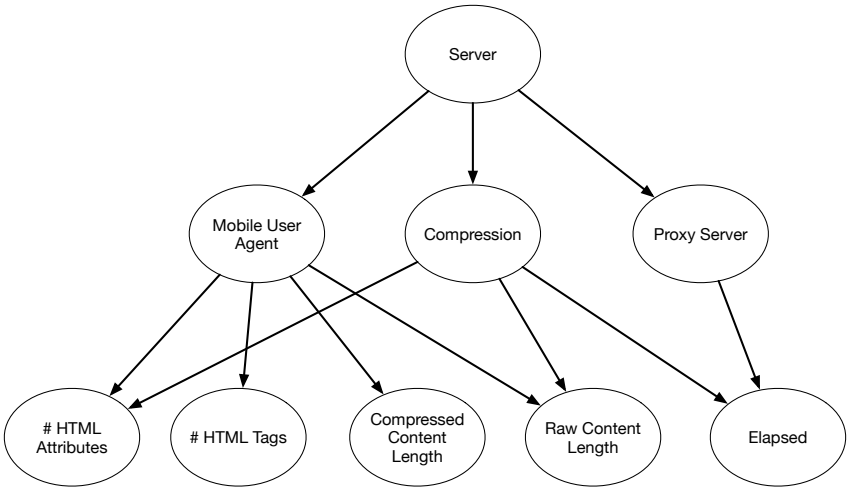

Figure B.3. Consistent model for the HTTP domain

B.1.3 Hypertext Transfer Protocol

For our experiment on HTTP & networking infrastructure, we used requests to specific web sites as subjects. We identified a number of target sites through a breadth-first web crawl initiated at `dmoz.org`. We ended the crawl after retrieving 5,472 sites. For 4,350 of those sites, we were able to issue successful web requests with all combinations treatments, yielding 34,800 subject-treatment combinations. We employed numerous techniques to ensure that content would not be cached, which could induce carryover across treatment regimes.

B.1.3.1 Treatments

- **Use of a Mobile User Agent**: Web browsers supply a *user agent* to identify themselves to the web servers that they request pages from. Some sites have different versions for mobile applications. We artificially adjusted the user agent from a standard user agent to a mobile user agent to explore this phenomena. This is accomplished with the HTTP `User-Agent` header.

- **Proxy Server**: Web requests can be routed through a *proxy*, a server which issues web requests on behalf of a client. The additional time required to route

the request to and from the proxy server can increase the elapsed time of the request. Our experiments were executed with Amazon EC2. Our "client" computers were making web requests from the east cost of the United States, and a proxy server was set up on the west coast.

- **Compression**: Applications can use the HTTP protocol to request that content be delivered with or without compression, possibly reducing the cross-network transmission time. In one compression configuration, the client requests `identity` compression, indicating that the content should be transmitted at face value. In another compression scheme, the client requests `gzip`, a common and effective scheme for HTTP content compression.

B.1.3.2 Outcomes

- **# of HTML Attributes and Tags**: These two outcomes describe the logical structure of the page. They may vary with respect to "mobile user agent".

- **Elapsed Time**: The time between issuance of the request and receipt of a response. This could be affected by network characteristics, which are determined in part by the time at which the request is issued and whether a proxy server is employed. Requests containing smaller payloads (influenced by compression) may also be faster to service.

- **Decompressed and Raw Content Length**: Two outcomes representing the size of a web page before and after content decompression, if applicable.

B.1.3.3 Subject Covariates

Only one subject covariate was identified for the HTTP domain, the web server reported via the `Server` header. This variable was coarsened into a version with 7 levels: Apache/2, Other Apache, Microsoft-IIS, nginx, Other, and Unknown.

B.2 Additional correlation matrices between causal modeling methods

Figure 5.9 presents correlation matrices between causal modeling algorithms, with correlations calculated across the 30 trials presented in Figure 5.2, for four representative data sets. Correlation results for all 37 data sets can be found in Figures B.4, B.5, and B.6.

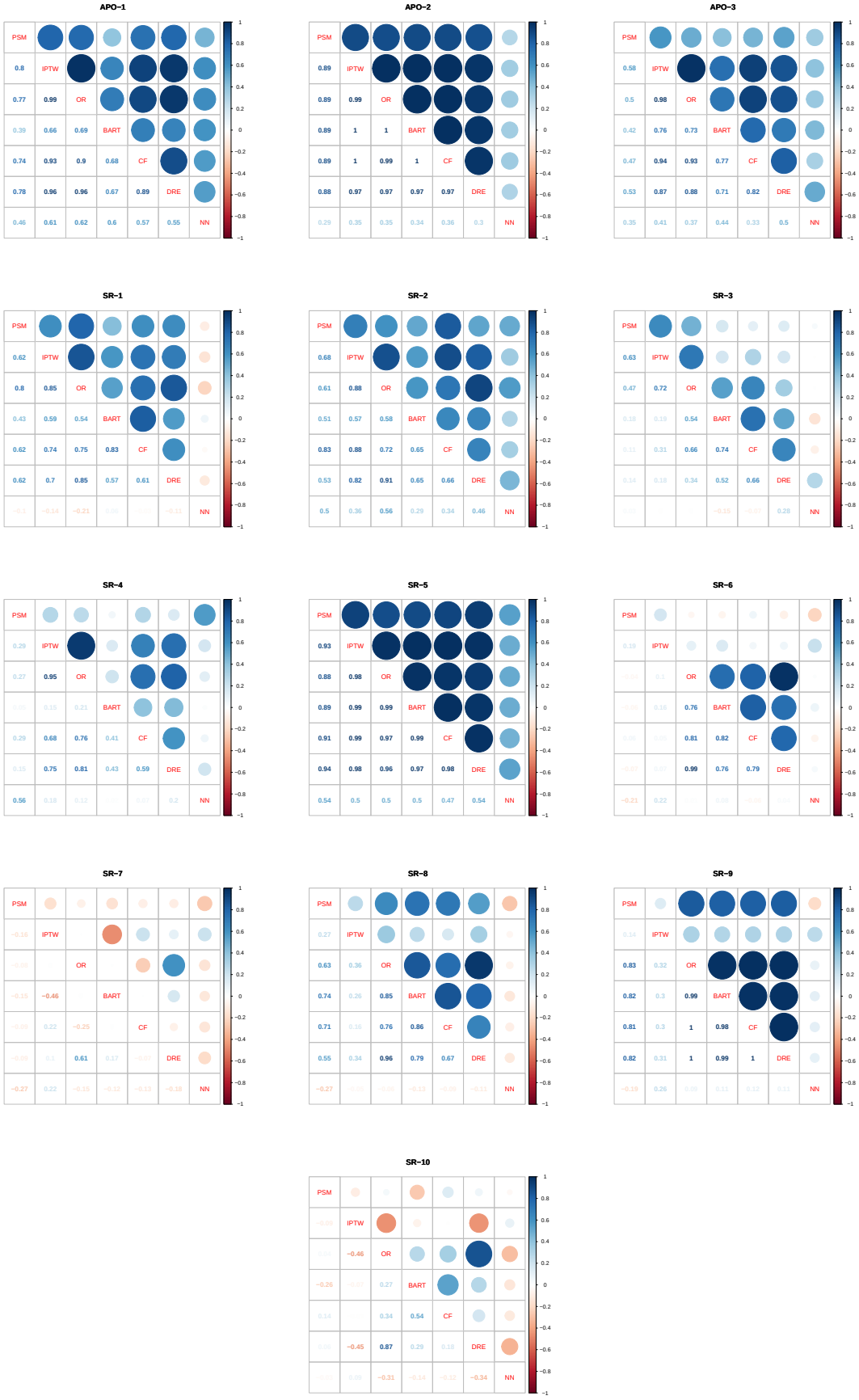

Figure B.4. Correlation matrices for APO 1-3 and SR 1-10

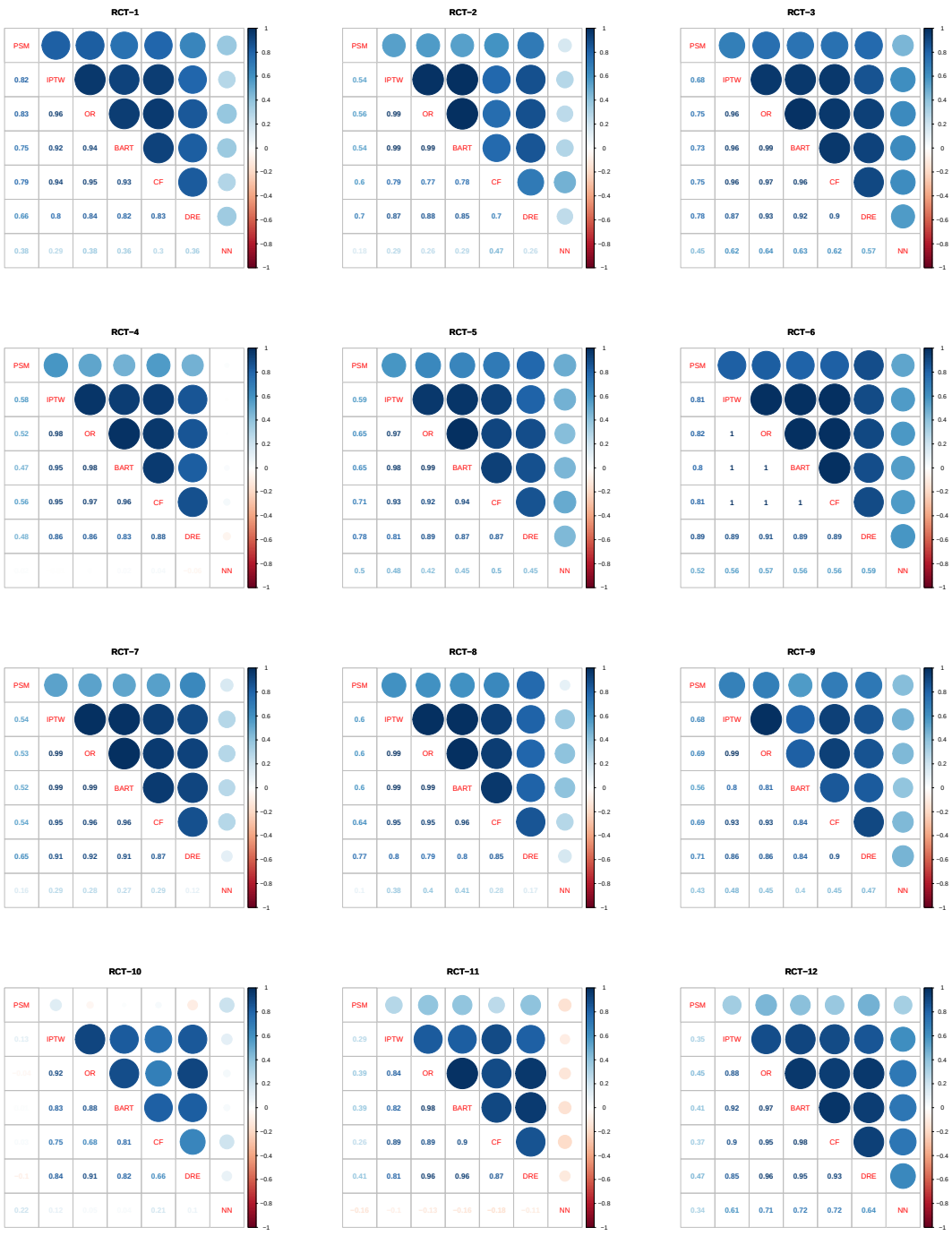

Figure B.5. Correlation matrices for RCT 1-12

Figure B.6. Correlation matrices for RCT 13-15 and Sim 1-9

BIBLIOGRAPHY

[1] Request for public comment on DRAFT NIH policy for data management and sharing and supplemental draft guidance, 2019.

[2] Data Observation Network for Earth, 2020. [Online; accessed 3-June-2020].

[3] Dryad, 2020. [Online; accessed 3-June-2020].

[4] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository, 2020. [Online; accessed 3-June-2020].

[5] NIH National Institute on Drug Abuse Data Share Website, 2020. [Online; accessed 3-June-2020].

[6] Project Data Sphere, 2020. [Online; accessed 3-June-2020].

[7] The Knowledge Network for Biocomplexity, 2020. [Online; accessed 3-June-2020].

[8] The National Institute of Mental Health Data Archive (NDA), 2020. [Online; accessed 3-June-2020].

[9] UK Data Service, 2020. [Online; accessed 3-June-2020].

[10] University of Michigan Institute for Social Research, 2020. [Online; accessed 3-June-2020].

[11] Vivli Center for Global Clinical Research Data, 2020. [Online; accessed 3-June-2020].

[12] Yale Institution for Social and Policy Studies Data Archive, 2020. [Online; accessed 3-June-2020].

[13] Achab, Massil, Bacry, Emmanuel, Gaïffas, Stéphane, Mastromatteo, Iacopo, and Muzy, Jean-François. Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research 18*, 1 (2017), 6998–7025.

[14] Acharya, Jayadev, Bhattacharyya, Arnab, Daskalakis, Constantinos, and Kandasamy, Saravanan. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems* (2018), pp. 9469–9481.

[15] Affeldt, Séverine, and Isambert, Hervé. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the 31st International Conference on Uncertainty in Artificial Intelligence* (2015), CEUR-WS. org, pp. 1–29.

[16] Agrawal, Raj, Broderick, Tamara, and Uhler, Caroline. Minimal i-map mcmc for scalable structure discovery in causal dag models. *arXiv preprint arXiv:1803.05554* (2018).

[17] Alrajeh, Dalal, Chockler, Hana, and Halpern, Joseph Y. Combining experts' causal judgments. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).

[18] Ambrogioni, Luca, Hinne, Max, Van Gerven, Marcel, and Maris, Eric. Gp cake: Effective brain connectivity with causal kernels. In *Advances in Neural Information Processing Systems* (2017), pp. 950–959.

[19] Arbour, David, Marazopoulou, Katerina, and Jensen, David. Inferring causal direction from relational data. In *Proceedings of the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016), AUAI Press, pp. 12–21.

[20] Arbour, David T, Marazopoulou, Katerina, Garant, Dan, and Jensen, David D. Propensity score matching for causal inference with relational data. In *Causality Workshop at the 30th International Conference on Uncertainty in Artificial Intelligence* (2014), pp. 25–34.

[21] Armen, Angelos P., and Evans, Robin J. Towards characterising bayesian network models under selection. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[22] Asbeh, Nuaman, and Lerner, Boaz. Pairwise cluster comparison for learning latent variable models. In *Causality Workshop at the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016).

[23] Athey, Susan, and Imbens, Guido. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*, 27 (2016), 7353–7360.

[24] Banzi, Rita, Canham, Steve, Kuchinke, Wolfgang, Krleza-Jeric, Karmela, Demotes-Mainard, Jacques, and Ohmann, Christian. Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials 20* (2019).

[25] Bareinboim, Elias, Forney, Andrew, and Pearl, Judea. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems* (2015), pp. 1342–1350.

[26] Bareinboim, Elias, and Tian, Jin. Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 3475–3481.

[27] Bareinboim, Elias, Tian, Jin, and Pearl, Judea. Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence* (2014), pp. 2410–2416.

[28] Bello, Kevin, and Honorio, Jean. Computationally and statistically efficient learning of causal bayes nets using path queries. In *Advances in Neural Information Processing Systems* (2018), pp. 10954–10964.

[29] Ben-Gal, Irad. Bayesian networks. *Encyclopedia of Statistics in Quality and Reliability 1* (2008).

[30] Ben-Michael, Eli, and Feller, Avi. Matrix constraints and multi-task learning for covariate balance. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[31] Blom, Tineke, Klimovskaia, Anna, Magliacane, Sara, and Mooij, Joris M. Causal discovery in the presence of measurement error. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 570–579.

[32] Blom, Tineke, and Mooij, Joris. Generalized structural causal models. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[33] Bongers, Stephan, and Mooij, Joris. Bridging the gap between random differential equations and structural causal models. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[34] Borboudakis, Giorgos, and Tsamardinos, Ioannis. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 1435–1444.

[35] Cai, Ruichu, Qiao, Jie, Zhang, Kun, Zhang, Zhenjie, and Hao, Zhifeng. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems* (2018), pp. 2671–2679.

[36] Cai, Ruichu, Qiao, Jie, Zhang, Zhenjie, and Hao, Zhifeng. Self: Structural equation likelihood framework for causal discovery. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).

[37] Chalupka, Krzysztof, Bischoff, Tobias, Perona, Pietro, and Eberhardt, Frederick. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016), AUAI Press, pp. 72–81.

[38] Chalupka, Krzysztof, Perona, Pietro, and Eberhardt, Frederick. Visual causal feature learning. In *Proceedings of the 31st International Conference on Uncertainty in Artificial Intelligence* (2015), AUAI Press, pp. 181–190.

[39] Chaudhry, Aditya, Xu, Pan, and Gu, Quanquan. Uncertainty assessment and false discovery rate control in high-dimensional granger causal inference. In *International Conference on Machine Learning* (2017), pp. 684–693.

[40] Cheng, Wei, Zhang, Kai, Chen, Haifeng, Jiang, Guofei, Chen, Zhengzhang, and Wang, Wei. Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 805–814.

[41] Chickering, David Maxwell. Optimal structure identification with greedy search. *Journal of Machine Learning Research 3* (Mar. 2003).

[42] Chikhaoui, Belkacem, Chiazzaro, Mauricio, and Wang, Shengrui. A new granger causal model for influence evolution in dynamic social networks: The case of dblp. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 51–57.

[43] Chipman, Hugh A, George, Edward I, and McCulloch, Robert E. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems* (2007), pp. 265–272.

[44] Cohen, Paul R. *Empirical Methods for Artificial Intelligence*, vol. 139. MIT press Cambridge, MA, 1995.

[45] Cook, Thomas D, and Campbell, Donald Thomas. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Rand McNally College Publishing Company Chicago, 1979.

[46] Cook, Thomas D., Shadish, William R., and Wong, Vivian C. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management 27*, 4 (2008), 724–750.

[47] Cooper, Gregory F, and Yoo, Changwon. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (1999), Morgan Kaufmann Publishers Inc., pp. 116–125.

[48] Correa, Juan D, and Bareinboim, Elias. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (2017), pp. 3740–3746.

[49] Cui, Ruifei, Groot, Perry, Schauer, Moritz, and Heskes, Tom. Learning the causal structure of copula models with latent variables. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 188–197.

[50] Dahlhaus, Rainer, and Eichler, Michael. Causality and graphical models in time series analysis. *Oxford Statistical Science Series* (2003), 115–137.

[51] Didelez, Vanessa. Causal reasoning for events in continuous time: a decision—theoretic approach. In *Proceedings of the 31st International Conference on Uncertainty in Artificial Intelligence* (2015), CEUR-WS. org, pp. 40–45.

[52] Dixit, Atray, Parnas, Oren, Li, Biyu, Chen, Jenny, Fulco, Charles P., Jerby-Arnon, Livnat, Marjanovic, Nemanja D., Dionne, Danielle, Burks, Tyler, Raychowdhury, Raktima, Adamson, Britt, Norman, Thomas M., Lander, Eric S., Weissman, Jonathan S., Friedman, Nir, and Regev, Aviv. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell 167*, 7 (2016), 1853–1866.e17.

[53] Dorie, Vincent, Hill, Jennifer, Shalit, Uri, Scott, Marc, and Cervone, Dan. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science 34* (2019), 43–68.

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

[55] Eckles, Dean, and Bakshy, Eytan. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint:1706.04692* (2017).

[56] Eckles, Dean, Kizilcec, René F., and Bakshy, Eytan. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences 113*, 27 (2016), 7316–7322.

[57] Elze, Markus C, Gregson, John, Baber, Usman, Williamson, Elizabeth, Sartori, Samantha, Mehran, Roxana, Nichols, Melissa, Stone, Gregg W, and Pocock, Stuart J. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology 69*, 3 (2017), 345–357.

[58] Figueiredo, Flavio, Borges, Guilherme Resende, de Melo, Pedro O. S. Vaz, and Assunção, Renato. Fast estimation of causal interactions using wold processes. In *Advances in Neural Information Processing Systems* (2018), pp. 2975–2986.

[59] Forré, Patrick, and Mooij, Joris M. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 269–278.

[60] Funk, Michele Jonsson, Westreich, Daniel, Wiesen, Chris, Stürmer, Til, Brookhart, M Alan, and Davidian, Marie. Doubly robust estimation of causal effects. *American Journal of Epidemiology 173*, 7 (2011), 761–767.

[61] Galles, David, and Pearl, Judea. Testing identifiability of causal effects. In *Proceedings of the 11th International Conference on Uncertainty in Artificial Intelligence* (1995), pp. 185–195.

[62] Gámez, José A, Mateo, Juan L, and Puerta, José M. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery 22*, 1-2 (2011), 106–148.

[63] Gao, Tian, and Ji, Qiang. Local causal discovery of direct causes and effects. In *Advances in Neural Information Processing Systems* (2015), pp. 2512–2520.

[64] Garant, Dan, and Jensen, David. Evaluating causal models by comparing interventional distributions. *arXiv preprint arXiv:1608.04698* (2016).

[65] Geiger, P., Janzing, D., and Schölkopf, B. Estimating causal effects by bounding confounding. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence* (Oregon, 2014), AUAI Press Corvallis, pp. 240–249.

[66] Geiger, Philipp, Zhang, Kun, Schoelkopf, Bernhard, Gong, Mingming, and Janzing, Dominik. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning* (2015), pp. 1917–1925.

[67] Gentzel, Amanda, Garant, Dan, and Jensen, David. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems 32*. 2019, pp. 11722–11732.

[68] Ghassami, AmirEmad, Kiyavash, Negar, Huang, Biwei, and Zhang, Kun. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems* (2018), pp. 6266–6276.

[69] Ghassami, AmirEmad, Kiyavash, Negar, Huang, Biwei, and Zhang, Kun. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems* (2018), pp. 6269–6279.

[70] Ghassami, AmirEmad, Salehkaleybar, Saber, Kiyavash, Negar, and Bareinboim, Elias. Budgeted experiment design for causal structure learning. *arXiv preprint arXiv:1709.03625* (2017).

[71] Ghassami, AmirEmad, Salehkaleybar, Saber, Kiyavash, Negar, and Zhang, Kun. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems* (2017), pp. 3011–3021.

[72] Glymour, Clark, Zhang, Kun, and Spirtes, Peter. Review of causal discovery methods based on graphical models. *Frontiers in Genetics 10* (2019), 524.

[73] Godlee, Fiona, and Groves, Trish. The new BMJ policy on sharing data from drug and device trials. *British Medical Journal 345* (2012).

[74] Gong, Mingming, Zhang, Kun, Schoelkopf, Bernhard, Tao, Dacheng, and Geiger, Philipp. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning* (2015), pp. 1898–1906.

[75] Gong, Mingming, Zhang, Kun, Schölkopf, Bernhard, Glymour, Clark, and Tao, Dacheng. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd International Conference on Uncertainty in Artificial Intelligence* (2017), vol. 2017, NIH Public Access.

[76] Gordon, Brett R, Zettelmeyer, Florian, Bhargava, Neha, and Chapsky, Dan. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science 38*, 2 (2019), 193–225.

[77] Granger, Clive WJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[78] Greenwald, Anthony G. Within-subjects designs: To use or not to use? *Psychological Bulletin 83*, 2 (1976), 314.

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

[80] Guyon, Isabelle, Aliferis, Constantin, Cooper, Greg, and Spirtes, Peter. Design and analysis of the Causation and Prediction Challenge. *WCCI 2008 Workshop on Causality* (2008), 1–33.

[81] Guyon, Isabelle, Janzing, Dominik, and Schölkopf, Bernhard. Causality: Objectives and assessment. *NIPS 2008 Workshop on Causality 6* (2010), 1–38.

[82] Hahn, P. Richard, Dorie, Vincent, and Murray, Jared S. Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017.

[83] Hahn, P Richard, Murray, Jared, and Carvalho, Carlos M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Confounding, and Heterogeneous Effects (October 5, 2017)* (2017).

[84] Hashimoto, Chikara, Torisawa, Kentaro, Kloetzer, Julien, and Oh, Jong-Hoon. Generating event causality hypotheses through semantic relations. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 2396–2403.

[85] Hensley-Alford, Sharon, Madan, Piyush, Mahatma, Shilpa, Italo Buleje, Yanyan Han, and Lu, Fang. Effect of secular trend in drug effectiveness study in real world data. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[86] Hernan, MA, and Robins, JM. *Causal Inference*. Chapman and HallCRC, forthcoming, 2020.

[87] Hill, Daniel N, Moakler, Robert, Hubbard, Alan E, Tsemekhman, Vadim, Provost, Foster, and Tsemekhman, Kiril. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 1839–1847.

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

[89] Hill, Jennifer L., Reiter, Jerome P., and Zanutoo, Elaine L. A comparison of experimental and observational data analyses. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (2004).

[90] Hoyer, Patrik O, Janzing, Dominik, Mooij, Joris M, Peters, Jonas, and Schölkopf, Bernhard. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* (2009), pp. 689–696.

[91] Hu, Huining, Li, Zhentao, and Vetta, Adrian R. Randomized experimental design for causal graph discovery. In *Advances in Neural Information Processing Systems* (2014), pp. 2339–2347.

[92] Hu, Shoubo, Chen, Zhitang, Nia, Vahid Partovi, Chan, Lai-Wan, and Geng, Yanhui. Causal inference and mechanism clustering of A mixture of additive noise models. In *Advances in Neural Information Processing Systems* (2018), pp. 5212–5222.

[93] Huang, Biwei, Zhang, Kun, Lin, Yizhu, Schölkopf, Bernhard, and Glymour, Clark. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018), ACM, pp. 1551–1560.

[94] Hyttinen, Antti, Eberhardt, Frederick, and Järvisalo, Matti. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence* (2014), pp. 340–349.

[95] Irfan, Mohammad T, and Ortiz, Luis E. Causal strategic inference in networked microfinance economies. In *Advances in Neural Information Processing Systems* (2014), pp. 1161–1169.

[96] Jaber, Amin, Zhang, Jiji, and Bareinboim, Elias. Causal identification under markov equivalence. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 978–987.

[97] Janzing, Dominik, and Schoelkopf, Bernhard. Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810* (2018).

[98] Javidian, Mohammad, and Valtorta, Marco. Finding minimal separators in ancestral graphs. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[99] Johansson, Fredrik, Shalit, Uri, and Sontag, David. Learning representations for counterfactual inference. In *International Conference on Machine Learning* (2016), PMLR, pp. 3020–3029.

[100] Kallus, Nathan. Causal inference by minimizing the dual norm of bias: Kernel matching and weighting estimators for causal effects. In *Causality Workshop at the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016).

[101] Kallus, Nathan, Mao, Xiaojie, and Udell, Madeleine. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems 31*. 2018, pp. 6921–6932.

[102] Kallus, Nathan, Mao, Xiaojie, and Udell, Madeleine. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems* (2018), pp. 6921–6932.

[103] Kallus, Nathan, Puli, Aahlad Manas, and Shalit, Uri. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems* (2018), pp. 10888–10897.

[104] Kallus, Nathan, and Zhou, Angela. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems* (2018), pp. 9269–9279.

[105] Kansky, Ken, Silver, Tom, Mély, David A, Eldawy, Mohamed, Lázaro-Gredilla, Miguel, Lou, Xinghua, Dorfman, Nimrod, Sidor, Szymon, Phoenix, Scott, and George, Dileep. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning* (2017), pp. 1809–1818.

[106] Keane, Michael P., and Wolpin, Kenneth I. Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior. *International Economic Review 48*, 4 (2007), 1351–1378.

[107] Kilbertus, Niki, Carulla, Mateo Rojas, Parascandolo, Giambattista, Hardt, Moritz, Janzing, Dominik, and Schölkopf, Bernhard. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* (2017), pp. 656–666.

[108] King, Gary, and Nielsen, Richard. Why propensity scores should not be used for matching. *Political Analysis 27*, 4 (2019).

[109] Kocaoglu, Murat, Dimakis, Alexandros G, Vishwanath, Sriram, and Hassibi, Babak. Entropic causal inference. In *Proceedings of the 31nd AAAI Conference on Artificial Intelligence* (2017), pp. 1156–1162.

[110] Kocaoglu, Murat, Shanmugam, Karthikeyan, and Bareinboim, Elias. Experimental design for latent variables. In *Advances in Neural Information Processing Systems* (2017), pp. 7018–7028.

[111] Kpotufe, Samory, Sgouritsa, Eleni, Janzing, Dominik, and Schölkopf, Bernhard. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning* (2014), pp. 478–486.

[112] Kruengkrai, Canasai, Torisawa, Kentaro, Hashimoto, Chikara, Kloetzer, Julien, Oh, Jong-Hoon, and Tanaka, Masahiro. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (2017), pp. 3466–3473.

[113] Kummerfeld, Erich, and Ramsey, Joseph. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 1655–1664.

[114] Kuntz, Richard E, Antman, Elliott M, Califf, Robert M, Ingelfinger, Julie R, Krumholz, Harlan M, Ommaya, Alexander, Peterson, Eric D, Ross, Joseph S, Waldstreicher, Joanne, Wang, Shirley V, et al. Individual patient-level data sharing for continuous learning: A strategy for trial data sharing. *NAM Perspectives* (2019).

[115] Kurutach, Thanard, Tamar, Aviv, Yang, Ge, Russell, Stuart J., and Abbeel, Pieter. Learning plannable representations with causal infogan. In *Advances in Neural Information Processing Systems* (2018), pp. 8747–8758.

[116] LaLonde, Robert J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review 76*, 4 (1986), pp. 604–620.

[117] Langley, Pat. The changing science of machine learning. *Machine Learning 82*, 3 (Mar 2011), 275–279.

[118] Lattimore, Finnian, Lattimore, Tor, and Reid, Mark D. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems* (2016), pp. 1181–1189.

[119] Le, Thuc, Hoang, Tao, Li, Jiuyong, Liu, Lin, Liu, Huawen, and Hu, Shu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2016).

[120] Lee, Kyungjae, Choi, Sungjoon, and Oh, Songhwai. Maximum causal tsallis entropy imitation learning. In *Advances in Neural Information Processing Systems* (2018), pp. 4408–4418.

[121] Lee, Sanghack, and Bareinboim, Elias. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems* (2018), pp. 2573–2583.

[122] Lee, Sanghack, and Honavar, Vasant. A characterization of markov equivalence classes of relational causal models under path semantics. In *Proceedings of the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016), AUAI Press, pp. 387–396.

[123] Lee, Sanghack, and Honavar, Vasant. On learning causal models from relational data. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (2016), pp. 3263–3270.

[124] Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining* (2011), pp. 297–306.

[125] Lin, Jianhua. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory 37*, 1 (Jan 1991), 145–151.

[126] Lindgren, Erik M., Kocaoglu, Murat, Dimakis, Alexandros G., and Vishwanath, Sriram. Experimental design for cost-aware learning of causal graphs. In *Advances in Neural Information Processing Systems* (2018), pp. 5284–5294.

[127] Lopez-Paz, David, Muandet, Krikamol, Schölkopf, Bernhard, and Tolstikhin, Iliya. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning* (2015), pp. 1452–1461.

[128] Louizos, Christos, Shalit, Uri, Mooij, Joris M, Sontag, David, Zemel, Richard, and Welling, Max. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (2017), pp. 6446–6456.

[129] Magliacane, Sara, Claassen, Tom, and Mooij, Joris M. Ancestral causal inference. In *Advances in Neural Information Processing Systems* (2016), pp. 4466–4474.

[130] Magliacane, Sara, van Ommen, Thijs, Claassen, Tom, Bongers, Stephan, Versteeg, Philip, and Mooij, Joris M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems* (2018), pp. 10846–10856.

[131] Magliacane, Sara, van Ommen, Thijs, Claassen, Tom, Bongers, Stephan, Versteeg, Philip, and Mooij, Joris M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems* (2018), pp. 10869–10879.

[132] Manjari Narayan, Molly Lucas, and Etkin, Amit. Learning time-varying bivariate causal structure using interventional neuroimaging. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[133] Marazopoulou, Katerina, Maier, Marc, and Jensen, David. Learning the structure of causal models with relational and temporal dependence. In *Proceedings of the 31st International Conference on Uncertainty in Artificial Intelligence* (2015), CEUR-WS. org, pp. 66–75.

[134] Margaritis, Dimitris, and Thrun, Sebastian. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems* (2000), pp. 505–511.

[135] Mary, Jérémie, Preux, Philippe, and Nicol, Olivier. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *International Conference on Machine Learning* (2014), pp. 172–180.

[136] McDonald, CJ, Hui, SL, and Tierney, WM. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing 9* (1992), 304–312.

[137] Meek, Christopher. Toward learning graphical and causal process models. In *Causality Workshop at the 30th International Conference on Uncertainty in Artificial Intelligence* (2014), p. 43.

[138] Merck, Christopher A, and Kleinberg, Samantha. Causal explanation under indeterminism: A sampling approach. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (2016), pp. 1037–1043.

[139] Miller, John, Hsu, Chloe, Troutman, Jordan, Perdomo, Juan, Zrnic, Tijana, Liu, Lydia, Sun, Yu, Schmidt, Ludwig, and Hardt, Moritz. Whynot, 2020.

[140] Mitrovic, Jovana, Sejdinovic, Dino, and Teh, Yee Whye. Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems* (2018), pp. 6986–6994.

[141] Mogensen, Søren Wengel, Malinsky, Daniel, and Hansen, Niels Richard. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 350–360.

[142] Mooij, Joris M., and Cremers, Jerome. An empirical study of one of the simplest causal prediction algorithms. In *Causality Workshop at the 31st International Conference on Uncertainty in Artificial Intelligence* (2015).

[143] Mooij, Joris M, Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob, and Schölkopf, Bernhard. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research 17*, 1 (2016), 1103–1204.

[144] Morgan, Stephen L, and Winship, Christopher. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[145] Nabi, Razieh, Kanki, Phyllis, and Shpitser, Ilya. Estimation of personalized effects associated with causal pathways. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 673–682.

[146] Nabi, Razieh, and Shpitser, Ilya. Semi-parametric causal sufficient dimension reduction of high dimensional treatment. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[147] Nie, Xinkun, and Wager, Stefan. Quasi-oracle estimation of heterogeneous treatment effects. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[148] Ogburn, Elizabeth L, Shpitser, Ilya, and Lee, Youjin. Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990* (2018).

[149] Ohmann, Christian, Banzi, Rita, Canham, Steve, Battaglia, Serena, Matei, Mihaela, Ariyo, Christopher, Becnel, Lauren, Bierer, Barbara, Bowers, Sarion, Clivio, Luca, Dias, Monica, Druml, Christiane, Faure, Hélène, Fenner, Martin, Galvez, Jose, Ghersi, Davina, Gluud, Christian, Groves, Trish, Houston, Paul, Karam, Ghassan, Kalra, Dipak, Knowles, Rachel L, Krleža-Jerić, Karmela, Kubiak, Christine, Kuchinke, Wolfgang, Kush, Rebecca, Lukkarinen, Ari, Marques, Pedro Silverio, Newbigging, Andrew, O'Callaghan, Jennifer, Ravaud, Philippe, Schlünder, Irene, Shanahan, Daniel, Sitter, Helmut, Spalding, Dylan, Tudur-Smith, Catrin, van Reusel, Peter, van Veen, Evert-Ben, Visser, Gerben Rienk, Wilson, Julia, and Demotes-Mainard, Jacques. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open 7*, 12 (2017).

[150] Pearl, Judea. Casual diagrams for empirical research. *Biometrika 82*, 4 (1995), 669–710.

[151] Pearl, Judea. *Causality*. Cambridge University Press, 2000.

[152] Peña, Jose M. Alternative markov and causal properties for acyclic directed mixed graphs. In *Proceedings of the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016), AUAI Press, pp. 577–586.

[153] Peters, J, and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural Computation 27*, 3 (March 2015), 771–799.

[154] Peña, Jose M. Identifiability of gaussian structural equation models with dependent errors having equal variances. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[155] Plis, Sergey, Danks, David, Freeman, Cynthia, and Calhoun, Vince. Rate-agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems* (2015), MIT Press, pp. 3303–3311.

[156] Pouliot, Guillaume. Modern methods for spatial econometrics. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[157] Razieh Nabi, Daniel Malinsky, and Shpitser, Ilya. Learning optimal fair policies. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

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

[159] Rothenhäusler, Dominik, Heinze, Christina, Peters, Jonas, and Meinshausen, Nicolai. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems* (2015), pp. 1513–1521.

[160] Roumpelaki, Anna, Borboudakis, Giorgos, Triantafillou, Sofia, and Tsamardinos, Ioannis. Marginal causal consistency in constraint-based causal learning. In *Causality Workshop at the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016).

[161] Rubenstein, Paul K., Bongers, Stephan, Mooij, Joris M., and Schölkopf, Bernhard. From deterministic odes to dynamic structural causal models. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 114–123.

[162] Rubenstein, Paul K, Tolstikhin, Ilya, Hennig, Philipp, and Schölkopf, Bernhard. Probabilistic active learning of functions in structural causal models. *arXiv preprint arXiv:1706.10234* (2017).

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

[164] Sachs, Karen, Perez, Omar, Pe'er, Dana, Lauffenburger, Douglas A, and Nolan, Garry P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*, 5721 (Apr. 2005), 523–529.

[165] Schaffter, Thomas, Marbach, Daniel, and Floreano, Dario. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics 27*, 16 (2011), 2263–2270.

[166] Schölkopf, B., Hogg, D., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of the 32nd International Conference on Machine Learning* (2015), vol. 37, JMLR, p. 2218–2226.

[167] Schwab, Patrick, Linhardt, Lorenz, and Karlen, Walter. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

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

[169] Shajarisales, Naji, Janzing, Dominik, Schölkopf, Bernhard, and Besserve, Michel. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning* (2015), pp. 285–294.

[170] Shalit, Uri, Johansson, Fredrik D, and Sontag, David. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning* (2017), PMLR, pp. 3076–3085.

[171] Shanmugam, Karthikeyan, Kocaoglu, Murat, Dimakis, Alexandros G, and Vishwanath, Sriram. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems* (2015), pp. 3195–3203.

[172] Sherman, Eli, and Shpitser, Ilya. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems* (2018), pp. 9446–9457.

[173] Shi, Claudia, Blei, David M, and Veitch, Victor. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120* (2019).

[174] Shimoni, Yishai, Yanover, Chen, Karavani, Ehud, and Goldschmnidt, Yaara. Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis. *arXiv preprint arXiv:1802.05046* (2018), 1–9.

[175] Shpitser, Ilya, and Sherman, Eli. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 530–539.

[176] Silva, Ricardo, and Evans, Robin. Causal inference through a witness protection program. In *Advances in Neural Information Processing Systems* (2014), pp. 298–306.

[177] Simmons, Joseph P, Nelson, Leif D, and Simonsohn, Uri. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*, 11 (2011), 1359–1366.

[178] Soleimani, Hossein, Subbaswamy, Adrash, and Saria, Suchi. Treatment-response models for counterfactual reasoning with countinuous-time, continuous-valued interventions. In *Proceedings of the 33rd International Conference on Uncertainty in Artificial Intelligence* (2017), vol. 2017, NIH Public Access.

[179] Spirtes, Peter, Glymour, Clark, and Scheines, Richard. *Causation, Prediction and Search*, 2nd ed. MIT Press, Cambridge, MA, 2000.

[180] Stanton, Andrew, Thart, Amanda, Jain, Ashish, Vyas, Priyank, Chatterjee, Arpan, and Shakarian, Paulo. Mining for causal relationships: A data-driven study of the islamic state. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 2137–2146.

[181] Steven Howard, Aaditya Ramdas, Jon McAuliffe, and Sekhon, Jasjeet. Uniform nonasymptotic confidence sequences for sequential treatment effect estimation. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. –.

[182] Subbaswamy, Adarsh, and Saria, Suchi. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *UAI* (2018), pp. 947–957.

[183] Subbaswamy, Adarsh, and Saria, Suchi. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 947–957.

[184] Sun, Wei, Wang, Pengyuan, Yin, Dawei, Yang, Jian, and Chang, Yi. Causal Inference via Sparse Additive Models with Application to Online Advertising. *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), 297–303.

[185] Sun, Wei, Wang, Pengyuan, Yin, Dawei, Yang, Jian, and Chang, Yi. Causal inference via sparse additive models with application to online advertising. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 297–303.

[186] Sun, Xiaohai, Janzing, Dominik, Schölkopf, Bernhard, and Fukumizu, Kenji. A kernel-based causal learning algorithm. In *Proceedings of the 24th International Conference on Machine Learning* (New York, NY, USA, 2007), ICML '07, ACM, pp. 855–862.

[187] Suvarna, Viraj Ramesh. Sharing individual patient data from clinical trials. *Perspectives in Clinical Research 6* (2015), 71–72.

[188] Toulis, Panagiotis, and Parkes, David C. Long-term causal effects via behavioral game theory. In *Advances in Neural Information Processing Systems* (2016), pp. 2604–2612.

[189] Triantafillou, Sofia. Score-based vs constraint-based causal learning in the presence of confounders. In *Causality Workshop at the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016).

[190] Tsamardinos, Ioannis, Aliferis, Constantin F, and Statnikov, Alexander R. Algorithms for large scale markov blanket discovery. In *FLAIRS* (2003), vol. 2.

[191] Tsamardinos, Ioannis, Brown, Laura E., and Aliferis, Constantin F. The max-min hill-climbing Bayesian network structure learning algorithm. *Journal of Machine Learning Research 65*, 1 (2006), 31–78.

[192] Tu, Ruibo, Zhang, Kun, Bertilson, Bo, Kjellstrom, Hedvig, and Zhang, Cheng. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In *Advances in Neural Information Processing Systems 32*. 2019, pp. 12793–12804.

[193] Wager, Stefan, and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* (2017).

[194] Wang, Ran. Identify heterogeneous effect and confounding effect via l_1-regularized soft decision tree. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[195] Wang, Yuhao, Solus, Liam, Yang, Karren, and Uhler, Caroline. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems* (2017), pp. 5822–5831.

[196] Wang, Yuhao, Squires, Chandler, Belyaeva, Anastasiya, and Uhler, Caroline. Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems* (2018), pp. 3770–3781.

[197] Wang, Yuhao, Squires, Chandler, Belyaeva, Anastasiya, and Uhler, Caroline. Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems* (2018), pp. 3774–3785.

[198] Wolfe, Elie, Spekkens, Robert W., and Fritz, Tobias. The inflation technique for causal inference with latent variables. In *Causality Workshop at the 34th International Conference on Uncertainty in Artificial Intelligence* (2018).

[199] Wu, Yongkai, Zhang, Lu, and Wu, Xintao. On discrimination discovery and removal in ranked data using causal graph. *arXiv preprint arXiv:1803.01901* (2018).

[200] Xu, Chang, Tao, Dacheng, and Xu, Chao. Large-margin multi-label causal feature learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 1924–1930.

[201] Xu, Hongteng, Farajtabar, Mehrdad, and Zha, Hongyuan. Learning granger causality for hawkes processes. In *International Conference on Machine Learning* (2016), pp. 1717–1726.

[202] Yabe, Akihiro, Hatano, Daisuke, Sumita, Hanna, Ito, Shinji, Kakimura, Naonori, Fukunaga, Takuro, and Kawarabayashi, Ken-ichi. Causal bandits with propagating inference. *arXiv preprint arXiv:1806.02252* (2018).

[203] Yang, Karren D, Katcoff, Abigail, and Uhler, Caroline. Characterizing and learning equivalence classes of causal dags under interventions. *arXiv preprint arXiv:1802.06310* (2018).

[204] Yeo, Jinyoung, Wang, Geungyu, Cho, Hyunsouk, Choi, Seungtaek, and Hwang, Seung-won. Machine-translated knowledge transfer for commonsense causal reasoning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).

[205] Yoo, Changwon, Thorsson, Vesteinn, and Cooper, Gregory F. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational dna microarray data. In *Biocomputing 2002*. World Scientific, 2001, pp. 498–509.

[206] Yoon, Jinsung, Jordon, James, and Van Der Schaar, Mihaela. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations* (2018).

[207] Zhalama, Zhang, Jiji, Eberhardt, Frederick, and Mayer, Wolfgang. Sat-based causal discovery under weaker assumptions. In *Proceedings of the 33rd International Conference on Uncertainty in Artificial Intelligence* (2017), vol. 2017, NIH Public Access.

[208] Zhang, Hao, Zhou, Shuigeng, and Guan, Jihong. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).

[209] Zhang, Hao, Zhou, Shuigeng, Zhang, Kun, and Guan, Jihong. Causal discovery using regression-based conditional independence tests. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (2017), pp. 1250–1256.

[210] Zhang, Junzhe, and Bareinboim, Elias. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems* (2018), pp. 3671–3681.

[211] Zhang, Junzhe, and Bareinboim, Elias. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems* (2018), pp. 3675–3685.

[212] Zhang, Junzhe, and Bareinboim, Elias. Fairness in decision-making–the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (2018).

[213] Zhang, Junzhe, and Bareinboim, Elias. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 653–662.

[214] Zhang, Kun, Gong, Mingming, Ramsey, Joseph, Batmanghelich, Kayhan, Spirtes, Peter, and Glymour, Clark. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768* (2017).

[215] Zhang, Kun, Gong, Mingming, Ramsey, Joseph, Batmanghelich, Kayhan, Spirtes, Peter, and Glymour, Clark. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence* (2018), pp. 1063–1072.

[216] Zhang, Kun, Gong, Mingming, and Schölkopf, Bernhard. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (2015), pp. 3150–3157.

[217] Zhang, Kun, Zhang, Jiji, Huang, Biwei, Schölkopf, Bernhard, and Glymour, Clark. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *Proceedings of the 32nd International Conference on Uncertainty in Artificial Intelligence* (2016), AUAI Press, pp. 825–834.

[218] Zhao, Qingyuan, Keele, Luke J., and Small, Dylan S. Comment: Will competition-winning methods for causal inference also succeed in practice? *Statistical Science 34*, 1 (02 2019), 72–76.

[219] Zhou, Yuxun, and Spanos, Costas J. Causal meets submodular: Subset selection with directed information. In *Advances in Neural Information Processing Systems* (2016), pp. 2649–2657.