A Survey on Causal Inference

LIUYI YAO, Alibaba Group
ZHIXUAN CHU and SHENG LI, University of Georgia
YALIANG LI, Alibaba Group
JING GAO, Purdue University
AIDONG ZHANG, University of Virginia

Causal inference is a critical research topic across many domains, such as statistics, computer science, education, public policy, and economics, for decades. Nowadays, estimating causal effect from observational data has become an appealing research direction owing to the large amount of available data and low budget requirement, compared with randomized controlled trials. Embraced with the rapidly developed machine learning area, various causal effect estimation methods for observational data have sprung up. In this survey, we provide a comprehensive review of causal inference methods under the potential outcome framework, one of the well-known causal inference frameworks. The methods are divided into two categories depending on whether they require all three assumptions of the potential outcome framework or not. For each category, both the traditional statistical methods and the recent machine learning enhanced methods are discussed and compared. The plausible applications of these methods are also presented, including the applications in advertising, recommendation, medicine, and so on. Moreover, the commonly used benchmark datasets as well as the open-source codes are also summarized, which facilitate researchers and practitioners to explore, evaluate and apply the causal inference methods.

CCS Concepts: • **Computing methodologies** → **Causal reasoning and diagnostics**; *Machine learning*; • **Information systems** → *Data mining*;

Additional Key Words and Phrases: Treatment effect estimation; Representation learning

ACM Reference format:
Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A Survey on Causal Inference. *ACM Trans. Knowl. Discov. Data* 15, 5, Article 74 (May 2021), 46 pages.
https://doi.org/10.1145/3444944

Work done when Liuyi Yao was a Ph.D. student at University at Buffalo.
This work is supported in part by the US National Science Foundation under grants IIS-1747614, IIS-2008208, IIS-1934600, IIS-1938167, and IIS-1955151.
Authors' addresses: L. Yao, Alibaba Group, 969 West Wen Yi Road, Yu Hang District, Hangzhou, Zhejiang, 311121, China; email: yly287738@alibaba-inc.com; Z. Chu, University of Georgia, 415 Boyd Graduate Studies Research Center, Athens, Georgia, 30602-7404, USA; email: zhixuan.chu@uga.edu; S. Li, University of Georgia, 415 Boyd Graduate Studies Research Center, Athens, Georgia, 30602-7404, USA; email: sheng.li@uga.edu; Y. Li, Alibaba Group, 500 108th Ave NE, Suite800, Bellevue, Washington, 98004, USA; email: yaliang.li@alibaba-inc.com; J. Gao, Purdue University, 465 Northwestern Ave., West Lafayette, Indiana, 47907-2035, USA; email: jinggao@purdue.edu; A. Zhang, University of Virginia, 85 Engineer's Way, Charlottesville, Virginia, 22904; email: aidong@virginia.edu.

1556-4681/2021/05-ART74 $15.00
https://doi.org/10.1145/3444944

1 INTRODUCTION

In everyday language, correlation and causality are commonly used interchangeably, although they have quite different interpretations. Correlation indicates a general relationship: two variables are correlated when they display an increasing or decreasing trend [7]. Causality is also referred to as cause and effect where the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. The main difference between causal inference and inference of correlation is that the former analyzes the response of the effect variable when the cause is changed [104, 150].

It is well known that "*correlation does not imply causation.*" For example, a study showed that girls have breakfast normally have lightweight than the girls who don't and thus concluded that having breakfast can help to lose weight. But in fact, these two events may just have correlation instead of causality. Maybe the girls who have breakfast every day have a better lifestyle, such as exercise frequently, sleep regularly, and have a healthy diet, which finally makes them have lightweight. In this case, having a better lifestyle is the common cause of both having breakfast and lightweight, and thus we also can treat it as a confounder of the causality between having breakfast and lightweight.

In many cases, it seems obvious that one action can cause another; however, there exists also many cases that we cannot easily tease out and make sure the relationship. Therefore, learning causality is one dauntingly challenging problem. The most effective way of inferring causality is to conduct a randomized controlled trial, which randomly assigns participants into a treatment group or a control group. As the randomized study is conducted, the only expected difference between the control and treatment groups is the outcome variable being studied. However, in reality, randomized controlled trials are always time-consuming and expensive, and thus the study cannot involve many subjects, which may be not representative of the real-world population a treatment/intervention would eventually target. Another issue is that the randomized controlled trials only focus on the average of samples, and it does not explain the mechanism or pertain for individual subjects. In addition, ethical issues also need to be considered in most of the randomized controlled trials, which largely limits its applications. Therefore, instead of the randomized controlled trials, the observational data is a tempting shortcut. Observational data is obtained by the researcher simply observing the subjects without any interfering. That means, the researchers have no control over treatments and subjects, and they just observe the subjects and record data based on their observations. From the observational data, we can find their actions, outcomes, and information about what has occurred, but cannot figure out the mechanism why they took a specific action. For the observational data, the core question is how to get the counterfactual outcome. For example, we want to answer this question "would this patient have different results if he received a different medication?" Answering such counterfactual questions is challenging due to two reasons [135]: the first one is that we only observe the factual outcome and never the counterfactual outcomes that would potentially have happened if they have chosen a different treatment option. The second one is that treatments are typically not assigned at random in observational data, which may lead the treated population differs significantly from the general population.

To solve these problems in causal inference from observational data, researchers develop various frameworks, including the potential outcome framework [127, 149] and the **structural causal model (SCM)** [102, 105, 107]. The potential outcome framework is also known as the Neyman–Rubin Potential Outcomes or the Rubin Causal Model. In the example, we mentioned above, a girl would have a particular weight if she had breakfast normally every day, whereas she would have a different weight if she didn't have breakfast normally. To measure the causal effect of

having breakfast normally for a girl, we need to compare the outcomes for the same person under both situations. Obviously, it is impossible to see both potential outcomes at the same time, and one of the potential outcomes is always missing. The potential outcome framework aims to estimate such potential outcomes and then calculate the treatment effect. Therefore, the treatment effect estimation is one of the central problems in causal inference under the potential outcome framework. Another influential framework in causal inference is the SCM, which includes the causal graph and the structural equations. The SCM describes the causal mechanisms of a system where a set of variables and the causal relationship among them are modeled by a set of simultaneous structural equations. Another line of learning causality is causal structure learning, whose objective is to reveal the causal relation by generating a causal graph. Representative methods can be divided into three categories, including constraint-based models [147], score-based models [31, 114], and functional causal models [62, 176]. Different from causal effect estimation, causal structure learning address a different class of problems, which is out of our survey's scope; see [148] for more information.

The causal inference has a close relationship with the machine learning area. In recent years, the magnificent bloom of the machine learning area enhances the development of the causal inference area. Powerful machine learning methods such as decision tree, ensemble methods, deep neural network, are applied to estimate the potential outcome more accurately. In addition to the amelioration of the outcome estimation model, machine learning methods also provide a new aspect to handle the confounders. Benefitting from the recently deep representation learning methods, the confounder variables are adjusted by learning the balanced representation for all covariates, so that conditioning on the learned representation, the treatment assignment is independent of the confounder variables. In machine learning, the more data the better. However, in causal inference, more data alone is not yet enough. Having more data only helps to get more precise estimates, but it cannot make sure these estimates are correct and unbiased. Machine learning methods enhance the development of causal inference, meanwhile, causal inference also helps machine learning methods. The simple pursuit of predictive accuracy is insufficient for modern machine learning research, and correctness and interpretability are also the targets of machine learning methods. Causal inference is starting to help to improve machine learning, such as recommender systems or reinforcement learning.

In this article, we provide a comprehensive review of the causal inference methods under the potential outcome framework. We first introduce the basic concepts of the potential outcome framework as well as its three critical assumptions to identify the causal effect. After that, various causal inference methods with these three assumptions are discussed in detail, including re-weighting methods, stratification methods, matching based methods, tree-based methods, representation-based methods, multi-task learning based methods, and meta-learning methods. Additionally, causal effect estimation methods that relax the three assumptions are also described to fulfill the needs in different settings. After introducing various causal effect estimation methods, the real-world applications that the discussed methods have great potential to benefit are discussed, including the advertisement area, recommendation area, medicine area, and reinforcement learning area as the representative examples.

To the best of our knowledge, this is the first article that provides a comprehensive survey for causal inference methods under the potential outcome framework. There also exist several surveys that discuss one category of the causal effect estimation methods, such as the survey of matching based methods [151], survey of tree-based and ensemble-based method [12], and the review of dynamic treatment regimes [28]. For the SCM, it is suggested to refer to the survey [104] or the book [103]. There is also a survey about learning causality from observational data [52] whose content ranges from inferring the causal graph from observational data, SCM, potential outcome

framework, and their connection to machine learning. Compared with the surveys mentioned above, this survey article mainly focuses on the theoretical background of the potential outcome framework, the representative methods across the statistic domain and machine learning domain, and how this framework and the machine learning area enhance each other.

To summarize, our contributions to this survey are as follows:

— *New taxonomy:* We separate various causal inference methods into two major categories based on whether they require the three assumptions of the potential outcome framework. The category requiring three assumptions are further divided into seven sub-categories based on the way to handle the confounder variables.
— *Comprehensive review:* We provide a comprehensive survey of the causal inference methods under the potential outcome framework. In each category, the detailed descriptions of the representative methods, the connection and comparison between the mentioned methods, and the general summation are provided.
— *Abundant resources:* In this survey, we list the state-of-art methods, the benchmark datasets, open-source codes, and representative applications.

The rest of the article is organized as follows. In Section 2, the background of the potential outcome framework is introduced, including the basic definitions, the assumptions, and the fundamental problems with their general solutions. In Section 3, the methods under three assumptions are presented. Then, in Section 4, we discuss the problem when some assumptions are not satisfied, and describe the methods that relax those assumptions. Next, we provide experimental guidelines in Section 5. Afterward, in Section 6, the typical applications of causal inference are illustrated. After that, in Section 7, the future directions and open problems are discussed. Finally, Section 8 summarizes the article.

2 BASIC OF CAUSAL INFERENCE

In this section, we present the background knowledge of causal inference, including task description, mathematical notions, assumptions, challenges, and general solutions. We also give an illustrative example that will be used throughout this survey.

Generally speaking, the task of causal inference is to estimate the outcome changes if another treatment had been applied. For example, suppose there are two treatments that can be applied to patients: Medicine A and Medicine B. When applying Medicine A to the interested patient cohort, the recovery rate is 70%, while applying Medicine B to the same cohort, the recovery rate is 90%. The change of recovery rate is the effect that treatment (i.e., medicine in this example) asserts on the recovery rate.

The above example describes an ideal situation to measure the treatment effect: applying different treatments to the same cohort. In real-world scenarios, this ideal situation can only be approximated by a randomized experiment, in which the treatment assignment is controlled, such as a completely random assignment. In this way, the group receives a specific treatment can be viewed as an approximation to the cohort we are interested in.

However, performing randomized experiments are expensive, time-consuming, and sometimes even unethical. Therefore, estimating the treatment effect from observational data has attracted growing attention due to the wide availability of observational data. Observational data usually contains a group of individuals taken different treatments, their corresponding outcomes, and possibly more information, but *without direct access to the reason/mechanism why they took the specific treatment*. Such observational data enable researchers to investigate the fundamental problem of learning the causal effect of a certain treatment without performing randomized experiments. To better introduce various treatment effect estimation methods, the following

section introduces several definitions, including unit, treatment, outcome, treatment effect, and other information (pre- and post-treatment variables) provided by observational data.

2.1 Definitions

Here we define the notations under the potential outcome framework [127, 149], which is logically equivalent to another framework, the SCM framework [72]. The foundation of the potential outcome framework is that the causality is tied to treatment (or action, manipulation, intervention), applied to a unit [69]. The treatment effect is obtained by comparing units' potential outcomes of treatments. In the following, we first introduce three essential concepts in causal inference: unit, treatment, and outcome.

Definition 1 (Unit). A unit is the atomic research object in the treatment effect study.

A unit can be a physical object, a firm, a patient, an individual person, or a collection of objects or persons, such as a classroom or a market, at a particular time point [69]. Under the potential outcome framework, the atomic research objects at different time points are different units. One unit in the dataset is a sample of the whole population, so in this survey, the term "sample" and "unit" are used interchangeably.

Definition 2 (Treatment). Treatment refers to the action that applies (exposes, or subjects) to a unit.

Let W ($W \in \{0, 1, 2, \ldots, N_W\}$) denote the treatment, where $N_W + 1$ is the total number of possible treatments. In the aforementioned medicine example, Medicine A is a treatment. Most of the literatures consider the binary treatment, and in this case, the group of units applied with treatment $W = 1$ is the *treated group*, and the group of units with $W = 0$ is the *control group*.

Definition 3 (Potential Outcome). For each unit-treatment pair, the outcome of that treatment when applied on that unit is the potential outcome [69].

The potential outcome of treatment with value w is denoted as $Y(W = w)$.

Definition 4 (Observed Outcome). The observed outcome is the outcome of the treatment that is actually applied.

The observed outcome is also called factual outcome, and we use Y^F to denote it where F stands for "factual." The relation between the potential outcome and the observed outcome is: $Y^F = Y(W = w)$ where w is the treatment actually applied.

Definition 5 (Counterfactual Outcome). Counterfactual outcome is the outcome if the unit had taken another treatment.

The counterfactual outcomes are the potential outcomes of the treatments except the one actually taken by the unit. Since a unit can only take one treatment, only one potential outcome can be observed, and the remaining unobserved potential outcomes are the counterfactual outcome. In the multiple treatment case, let $Y^{CF}(W = w')$ denote the counterfactual outcome of treatment with value w'. In the binary treatment case, for notation simplicity, we use Y^{CF} to denote the counterfactual outcome, and $Y^{CF} = Y(W = 1 - w)$, where w is the treatment actually taken by the unit.

In the observational data, besides the chosen treatments and the observed outcome, the units' other information is also recorded, and they can be separated as pre-treatment variables and the post-treatment variables.

Definition 6 (Pre-treatment Variables). Pre-treatment variables are the variables that are not affected by the treatment.

Pre-treatment variables are also named as *background variables*, and they can be patients' demographics, medical history, and and the like. Let X denote the pre-treatment variables.

Definition 7 (Post-treatment Variables). The post-treatment variables are the variables that are affected by the treatment.

One example of post-treatment variables is the intermediate outcome, such as the lab test after taking the medicine in the aforementioned medicine example.

In the following sections, the terminology *variable* refers to the pre-treatment variable unless otherwise specified.

Treatment Effect. After introducing the observational data and the key terminologies, the treatment effect can be quantitatively defined using the above definitions. The treatment effect can be measured at the population, treated group, subgroup, and individual levels. To make these definitions clear, here we define the treatment effect under binary treatment, and it can be extended to multiple treatments by comparing their potential outcomes.

At the population level, the treatment effect is named as the **Average Treatment Effect (ATE)**, which is defined as:

$$\text{ATE} = \mathbb{E}[\mathbf{Y}(W = 1) - \mathbf{Y}(W = 0)], \tag{1}$$

where $\mathbf{Y}(W = 1)$ and $\mathbf{Y}(W = 0)$ are the potential treated and control outcome of the whole population respectively.

For the treated group, the treatment effect is named as **Average Treatment effect on the Treated group (ATT)**, and it is defined as:

$$\text{ATT} = \mathbb{E}[\mathbf{Y}(W = 1)|W = 1] - \mathbb{E}[\mathbf{Y}(W = 0)|W = 1], \tag{2}$$

where $\mathbf{Y}(W = 1)|W = 1$ and $\mathbf{Y}(W = 0)|W = 1$ are the potential treated and control outcome of the treated group respectively.

At the subgroup level, the treatment effect is called **Conditional Average Treatment Effect (CATE)**, which is defined as:

$$\text{CATE} = \mathbb{E}[\mathbf{Y}(W = 1)|X = x] - \mathbb{E}[\mathbf{Y}(W = 0)|X = x], \tag{3}$$

where $\mathbf{Y}(W = 1)|X = x$ and $\mathbf{Y}(W = 0)|X = x$ are the potential treated and control outcome of the subgroup with $X = x$, respectively. CATE is a common treatment effect measurement under the case where the treatment effect varies across different subgroups, which is also known as the heterogeneous treatment effect.

At the individual level, the treatment effect is called **Individual Treatment Effect (ITE)**, and the ITE of unit i is defined as:

$$\text{ITE}_i = Y_i(W = 1) - Y_i(W = 0), \tag{4}$$

where $Y_i(W = 1)$ and $Y_i(W = 0)$ are the potential treated and control outcome of unit i, respectively. In some literatures [70, 139], the ITE is viewed equivalent to the CATE.

Objective. For causal inference, our objective is to estimate the treatment effects from the observational data. Formally speaking, given the observational dataset, $\{X_i, W_i, Y_i^F\}_{i=1}^N$, where N is the total number of units in the datasets, the goal of the causal inference task is to estimate the treatment effects defined above.

2.2 An Illustrative Example

To better illustrate causal inference, we use the following example combined with the notations defined above to give an overview. In this example, we want to evaluate the treatment effects of several different medications for one disease, by exploiting the observational data (i.e., the electronic health records) that include demographic information of patients, the specific medication with the specific dosage taken by patients, and the outcome of medical tests. Obviously, we can only get one factual outcome for a specific patient from electronic health records, and thus the core task is to predict what would have happened if a patient took another treatment (i.e., a different medication or the same medication with a different dosage). Answering such counterfactual questions is very challenging. Therefore, we want to use causal inference to predict all of the potential outcomes for each patient over all of the medications with different dosages. Then, we can reasonably and accurately evaluate and compare the treatment effect of different medications for this disease.

One particular point to keep in mind is that for each medication, they may have different dosages. For example, for medication A, the dosage range can be a continuous variable in the range $[a, b]$, while for medication B, the dosage can be a categorical variable that has several specific dosage regimens.

In the aforementioned example, the units are the patients with the studied disease. The treatments refer to the different medications with specific dosages for this disease, and we use W ($W \in \{0, 1, 2, \ldots, N_W\}$) to denote these treatments. For example, $W_i = 1$ can represent the medication A with a specific dosage is taken by the unit i, and $W_i = 2$ represents the medication B with a specific dosage is taken by the unit i. Y is the outcome, such as one type of blood test that can measure the medication's ability to destroy the disease and lead to the recovery of the patients. Let $Y_i(W = 1)$ denote the potential outcome of medication A with a specific dosage on patient i. The features of patients may include age, gender, clinical presentation, some other medical tests, and the like. Among these features, age, gender, and other demographic information are pre-treatment variables that cannot be affected by taking treatment. Some clinical presentations and medical tests are affected by taking medications, and they are post-treatment variables. In this example, our goal to estimate the treatment effects of different medications for this disease based on the provided observational data.

In the following sections, we will continuously use this example to explain more concepts and illustrate intuitions behind various causal inference methods.

2.3 Assumptions

In order to estimate the treatment effect, the following assumptions are commonly used in the causal inference literature.

Assumption 2.1 [Stable Unit Treatment Value Assumption (SUTVA)]. *The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

This assumption emphasizes two points: The first point is the independence of each unit, that is, there are no interactions between units. In the context of the above illustrative example, one patient's outcome will not affect other patients' outcomes.

The second point is the single version for each treatment. In the above example, Medicine A with different dosages are different treatments under the SUTVA assumption.

Assumption 2.2 [Ignorability]. *Given the background variable, X, treatment assignment W is independent to the potential outcomes, i.e., $W \perp\!\!\!\perp Y(W = 0), Y(W = 1)|X$.*

In the context of the illustrative example, this ignorability assumption indicates two folds: first, if two patients have the same background variable X, their potential outcomes should be the same whatever the treatment assignment is, i.e., $p(Y_i(0), Y_i(1)|X = x, W = W_i) = p(Y_j(0), Y_j(1)|X = x, W = W_j)$. Analogously, if two patients have the same background variable value, their treatment assignment mechanism should be same whatever the value of potential outcomes they have, i.e., $p(W|X = x, Y_i(0), Y_i(1)) = p(W|X = x, Y_j(0), Y_j(1))$. The ignorability assumption is also named as unconfoundedness assumption. With this unconfoundedness assumption, for the units with the same background variable X, their treatment assignment can be viewed as random.

Assumption 2.3 [Positivity]. *For any value of X, treatment assignment is not deterministic:*

$$P(W = w|X = x) > 0, \quad \forall w \text{ and } x. \tag{5}$$

If for some values of X, the treatment assignment is deterministic; then for these values, the outcomes of at least one treatment could never be observed. In this case, it would be unable and meaningless to estimate the treatment effect. To be more specific, suppose there are two treatments: Medicine A and Medicine B. Let's assume that patients with age greater than 60 are always assigned with medicine A, then it will be unable and meaningless to study the outcome of medicine B on those patients. In other words, the positivity assumption indicates the variability, which is important for treatment effect estimation.

In [69], the ignorability and the positivity assumptions together are called *Strong Ignorability* or *Strongly Ignorable Treatment Assignment*.

With these assumptions, the relationship between the observed outcome and the potential outcome can be rewritten as:

$$\begin{aligned} \mathbb{E}[Y(W = w)|X = x] &= \mathbb{E}[Y(W = w)|W = w, X = x] \text{ (Ignorability)} \\ &= \mathbb{E}[Y^F|W = w, X = x], \end{aligned} \tag{6}$$

where Y^F is the random variable of the observed outcome, and $Y(W = w)$ is the random variable of the potential outcome of treatment w. If we are interested in the potential outcome of one specific group (either the subgroup, the treated group, or the whole population), the potential outcome can be obtained by taking expectation of the observed outcome over that group.

With the above equation, we can rewrite the treatment effect defined in Section 2.1 as follows:

$$\begin{aligned} \text{ITE}_i &= W_i Y_i^F - W_i Y_i^{CF} + (1 - W_i) Y_i^{CF} - (1 - W_i) Y_i^F \\ \text{ATE} &= \mathbb{E}_X \left[\mathbb{E}[Y^F|W = 1, X = x] - \mathbb{E}[Y^F|W = 0, X = x] \right] \\ &= \frac{1}{N} \sum_i (Y_i(W = 1) - Y_i(W = 0)) = \frac{1}{N} \sum_i \text{ITE}_i \\ \text{ATT} &= \mathbb{E}_{X_T} \left[\mathbb{E}[Y^F|W = 1, X = x] - \mathbb{E}[Y^F|W = 0, X = x] \right] \\ &= \frac{1}{N_T} \sum_{\{i:W_i=1\}} (Y_i(W = 1) - Y_i(W = 0)) = \frac{1}{N_T} \sum_{\{i:W_i=1\}} \text{ITE}_i \\ \text{CATE} &= \mathbb{E}[Y^F|W = 1, X = x] - \mathbb{E}[Y^F|W = 0, X = x] \\ &= \frac{1}{N_x} \sum_{\{i:X_i=x\}} (Y_i(W = 1) - Y_i(W = 0)) = \frac{1}{N_x} \sum_{\{i:X_i=x\}} \text{ITE}_i \end{aligned} \tag{7}$$

where $Y_i(W = 1)$ and $Y_i(W = 0)$ are the potential treated/control outcomes of unit i, N is the total number of units in the whole population, N_T is the number of units in the treated group, and N_x is the number of units in the group with $X = x$. The second line in the ATE, ATT, and CATE equations are their empirical estimations. Empirically, the ATE can be estimated as the average of

Table 1. An Example to Show the Spurious Effect of Confounder Variable *Age*

Recovery Rate / Treatment / Age	Treatment A	Treatment B
Young	234/270 = 87%	81/87 = **92**%
Older	55/80 = 69%	192/263 = **73**%
Overall	289/350 = **83**%	273/350 = 78%

ITE on the entire population. Similarly, ATT and CATE can be estimated as the average of ITE on the treated group and specific subgroup separately.

However, due to the fact that the potential treated/control outcomes can never be observed simultaneously, the key point in the treatment effect estimation is how to estimate the counterfactual outcome in ITE estimation or how to estimate the $\frac{1}{N_*}\sum_i Y_i(W=1)$ and $\frac{1}{N_*}\sum_i Y_i(W=0)$, where N_* denotes N, N_T or N_x. In the following section, we will discuss the challenges in estimation these terms and briefly introduce the general solutions.

2.4 Confounders and General Solutions

As mentioned above, how to estimate the average potential treated/control outcome over a specific group is the core of causal inference. Let's take ATE as a case study: when estimating the ATE, a natural idea is to directly use the average of observed treated/control outcomes, i.e., $\hat{\text{ATE}} = \frac{1}{N_T}\sum_{i=1}^{N_T} Y_i^F - \frac{1}{N_C}\sum_{i=1}^{N_C} Y_j^F$, where N_T and N_C is the number of units in the treated and control group, respectively. However, due to the existence of *confounders*, there is a serious problem in this estimation: this calculated ATE includes a spurious effect brought by the confounders.

Definition 8 (Confounders). Confounders are the variables that affect both the treatment assignment and the outcome.

Confounders are some special pre-treatment variables, such as age in the medicine example. When directly using the average of observed treated/control outcome, the calculated ATE includes not only the effect of treatment on the outcome but also the effect of confounders on the outcome, which leads to the *spurious effect*. For example, in the medicine example, age is a confounder. Age affects the recovery rate: in general, young patients have a better chance to recover compared to older patients. Age also affects the treatment choice: young patients may prefer to take medicine A while older patients prefer medicine B, or for the same medicine, young patients have a different dosage with elderly patients. The observational data is shown in Table 1, and let us estimate ATE according to the above equation: $\hat{\text{ATE}} = \frac{1}{N_A}\sum_{i=1}^{N_A} Y_i^F - \frac{1}{N_B}\sum_{i=1}^{N_B} Y_j^F = 289/350 - 273/350 = 5\%$, where N_A and N_B is the number of patients taking Medicine A and B, respectively. However, we cannot conclude that Treatment A is more effective than Treatment B, because the high average recovery rate of the group taking Treatment A may be caused by the fact that most patients of this group (270 out of 350) are young patients. Thus the effect of age on the recovery rate is the spurious effect, as it is mistakenly counted into the effect of treatment on the outcome.

From Table 1, we can observe another interesting phenomenon, *Simpson's paradox* (or Simpson's reversal, Yule–Simpson effect, amalgamation paradox, reversal paradox) [21, 50], brought by the confounder. It can be observed that: in both Young and Older patient groups, Medicine B has a higher recovery rate than Medicine A; but when combining these two groups, Medicine A is the one with a higher recovery rate. This paradox is caused by the confounder variable: when compare the recovery rate in the whole group, most of the people taking medicine A are young, and the comparison shown in the table fails to eliminate the effect of age on the recovery rate.

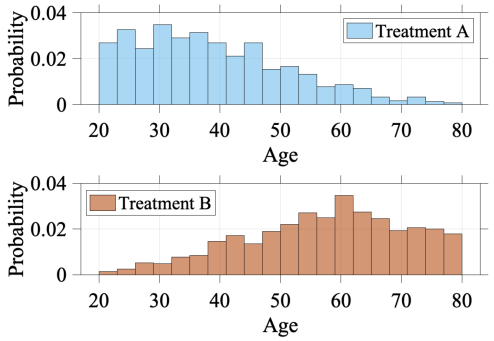

Fig. 1. An example to show the selection bias caused by confounder variable *Age*.

In addition to the spurious effect in treatment effect estimation, confounders also cause problems in counterfactual outcome estimation. As shown in Equation (7), counterfactual outcome estimation is an alternative way to estimate the ATE. Confounder variables cause selection bias, which makes the counterfactual outcome estimation more difficult.

Selection bias is the phenomenon that the distribution of the observed group is not representative to the group we are interested in, i.e., $p(X_{obs}) \neq p(X_*)$, where $p(X_{obs})$ and $p(X_*)$ are the distributions of the variables in the observed group and the interested group, respectively. Confounder variables affect units' treatment choices, which leads to selection bias. In the medicine example, age is a confounder variable, so that people of different ages have different treatment preferences. Figure 1 shows the age distribution of the observed treated/control group. Apparently, the age distribution of the observed treated group is different from the age distribution of the observed control group. This phenomenon exacerbates the difficulty of counterfactual outcome estimation as we need to estimate the control outcome of units in the treated group based on the observed control group, and similarly, estimate the treated outcome of units in the control group based on the observed treated group. If we directly train the potential outcome estimation model $\hat{Y}(x, w) = f_w(x)$ on the data with $W = w$ without handling the selection bias, the trained model would work poorly in estimating the potential outcome of $W = w$ for the units in the other group. This problem brought by the selection is also named as covariate shift in the Machine Learning community.

Handing the problems caused by confounder variables is the essential part of causal inference, and the procedure of handing confounder variables is called *adjust confounders*. The following part of this section briefly discusses the general solutions to tackle the above two problems caused by confounders under the ignorability assumption. The problem when there exists unobserved confounders will be discussed in Section 4.2.

To solve the spurious effect problem, we should take the effect of confounder variables on outcomes into consideration. A general approach along this direction first estimates the treatment effect conditioning on the confounder variables and then conducts weighted averaging over the confounder according to its distribution. To be more specific,

$$\begin{aligned}\hat{\text{ATE}} &= \textstyle\sum_x p(x)\mathbb{E}[Y^F|X = x, W = 1] - \sum_x p(x)\mathbb{E}[Y^F|X = x, W = 0] \\ &= \textstyle\sum_{\mathcal{X}^*} p(X \in \mathcal{X}^*) \left(\frac{1}{N_{\{i:X_i \in \mathcal{X}^*, W_i = 1\}}} \sum_{\{i:X_i \in \mathcal{X}^*, W_i = 1\}} Y_i^F \right) \\ &\quad - \textstyle\sum_{\mathcal{X}^*} p(X \in \mathcal{X}^*) \left(\frac{1}{N_{\{j:X_j \in \mathcal{X}^*, W_j = 1\}}} \sum_{\{j:X_j \in \mathcal{X}^*, W_j = 0\}} Y_j^F \right),\end{aligned} \tag{8}$$

where $\mathcal{X}^*$ is a set of X values, $p(X \in \mathcal{X}^*)$ is the probability of the background variables in $\mathcal{X}^*$ over the whole population, $\{i : x_i \in \mathcal{X}^*, W_i = w\}$ is the subgroup of units whose background variable values belong to $\mathcal{X}^*$ and treatment is equal to w. Stratification, which will be discussed in details later, is a representative method of this category.

For the selection bias problem, there are two general approaches to solve it. The first general approach handles selection bias by creating a pseudo group which is approximately close to the interested group. Possible methods include sample re-weighting, matching, tree-based methods, confounder balancing, balanced representation learning methods, multi-task-based methods, and the like. The created pseudo-group alleviates the negative influence of the selection bias, and better counterfactual outcome estimations can be obtained. The other general approach first trains the base potential outcome estimation models solely on the observed data, and then correct the estimation bias caused by the selection bias. Meta-learning based methods belong to this category.

3 CAUSAL INFERENCE METHODS RELYING ON THREE ASSUMPTIONS

In this section, we introduce existing causal inference methods that rely on the three assumptions introduced in Section 2. According to the way to control confounders, we divide these methods into the following categories: (1) re-weighting methods; (2) stratification methods; (3) matching methods; (4) tree-based methods; (5) representation based methods; (6) multi-task methods; and (7) meta-learning methods.

3.1 Re-weighting Methods

Due to the existence of confounders, the covariate distributions of the treated group and control group are different, which leads to the *selection bias* problem as described in Section 2.4. In other words, the treatment assignment is correlated with covariates in the observational data. Sample re-weighting is an effective approach to overcome selection bias. By assigning appropriate weight to each unit in the observational data, a pseudo-population can be created on which the distributions of the treated group and control group are similar.

In sample re-weighting methods, a key concept is *balancing score*. Balancing score $b(x)$ is a general weighting score, which is the function of x satisfying: $W \perp\!\!\!\perp |b(x)$ [69], where W is the treatment assignment and x is the background variables. There are various designs of the balancing score, and apparently, the most trivial design of balancing score is $b(x) = x$ due to the ignorability assumption. Besides, propensity score is also a special case of balancing score.

Definition 9 (Propensity Score). The propensity score is defined as the conditional probability of treatment given background variables [122]:

$$e(x) = Pr(W = 1|X = x). \tag{9}$$

In detail, a propensity score indicates the probability of a unit being assigned to a particular treatment given a set of observed covariates. Balancing scores that incorporate propensity score are the most common approach.

A summarization of the algorithms mentioned in this section is shown in Figure 2. The propensity score based sample reweighting will be introduced in the next section, followed by methods that weigh both samples and the covariates.

3.1.1 Propensity Score Based Sample Re-weighting. Propensity scores can be used to reduce selection bias by equating groups based on these covariates. **Inverse propensity weighting (IPW)** [121, 122], also named as inverse probability of treatment weighting, assigns a weight r to each sample:

$$r = \frac{W}{e(x)} + \frac{1-W}{1-e(x)}, \tag{10}$$

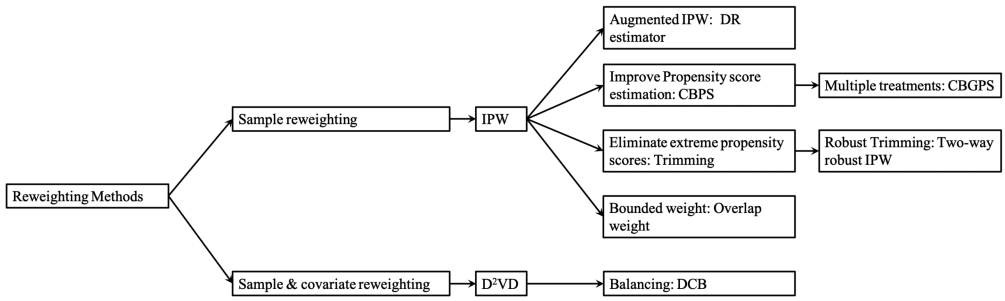

Fig. 2. Categorization of Reweighting Methods.

where W is the treatment assignment ($W = 1$ denotes being treated group; $W = 0$ denotes the control group), and $e(x)$ is the propensity score defined in Equation (9).

After re-weighting, the IPW estimator of ATE is:

$$\hat{\text{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i^F}{1-\hat{e}(x_i)}, \tag{11}$$

and its normalized version, which is preferred especially when the propensity scores are obtained by estimation [68]:

$$\hat{\text{ATE}}_{IPW} = \sum_{i=1}^{n}\frac{W_i Y_i^F}{\hat{e}(x_i)}\Big/\sum_{i=1}^{n}\frac{W_i}{\hat{e}(x_i)} - \sum_{i=1}^{n}\frac{(1-W_i)Y_i^F}{1-\hat{e}(x_i)}\Big/\sum_{i=1}^{n}\frac{(1-W_i)}{1-\hat{e}(x_i)}. \tag{12}$$

Both large and small sample theory show that adjustment for the scalar propensity score is enough to remove bias due to all observed covariates [122]. The propensity score can be used to balance the covariates in the treatment and control groups and therefore reduce the bias through matching, stratification (subclassification), regression adjustment, or some combination of all three. [37] discusses the use of propensity score to reduce the bias, which also provides examples and detailed discussions.

However, in practice, the correctness of the IPW estimator highly relies on the correctness of the propensity score estimation, and slightly misspecification of propensity scores would cause ATE estimation error dramatically [66]. To handle this dilemma, **Doubly Robust (DR)** estimator [119], also named as Augmented IPW, is proposed. DR estimator combines the propensity score weighting with the outcome regression, so that the estimator is robust even when one of the propensity score or outcome regression is incorrect (but not both). In detail, the DR estimator is formalized as:

$$\begin{aligned}\hat{\text{ATE}}_{DR} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{W_i Y_i^F}{\hat{e}(x_i)} - \frac{W_i-\hat{e}(x_i)}{\hat{e}(x_i)}\hat{m}(1,x_i)\right] - \left[\frac{(1-W_i)Y_i^F}{1-\hat{e}(x_i)} - \frac{W_i-\hat{e}(x_i)}{1-\hat{e}(x_i)}\hat{m}(0,x_i)\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}(1,x_i) + \frac{W_i(Y_i^F-\hat{m}(1,x_i))}{\hat{e}(x_i)} - \hat{m}(0,x_i) - \frac{(1-W_i)(Y_i^F-\hat{m}(0,x_i))}{1-\hat{e}(x_i)}\right\},\end{aligned} \tag{13}$$

where $\hat{m}(1,x_i)$ and $\hat{m}(0,x_i)$ are the regression model estimations of treated and control outcomes. The DR estimator is consistent and therefore asymptotically unbiased, if either the propensity score is correct or the model correctly reflects the true relationship among exposure and confounders with the outcome [46]. In reality, one definitely cannot guarantee whether one model can accurately explain the relationship among variables. The combination of outcome regression

with weighting by propensity score ensures that the estimators are robust to misspecification of one of these models [14, 117, 119, 133].

The DR estimator consults outcomes to make the IPW estimator robust when propensity score estimation is not correct. An alternative way is to improve the estimation of propensity scores. In the IPW estimator, propensity score serves as both the probability of being treated and the covariate balancing score, **covariate balancing propensity score (CBPS)** [66] is proposed to exploit such dual characteristics. In particular, CBPS estimates propensity scores by solving the following problem:

$$\mathbb{E}\left[\frac{W_i\tilde{x_i}}{e(x_i;\beta)} - \frac{(1-W_i)\tilde{x_i}}{1-e(x_i;\beta)}\right] = 0, \tag{14}$$

where $\tilde{x_i} = f(x_i)$ is a predefined vector-valued measurable function of x_i. By solving the above problem, CBPS directly constructs the covariate balancing score from the estimated parametric propensity score, which increase the robustness to the misspecification of the propensity score model. An extension of CBPS is the **covariate balancing generalized propensity score (CBGPS)** [47], which enables to handle the treatment with continuous value. Due to the continuous valued treatment, it's hard to directly minimized the covariates distribution distance between the control and treated group. CBGPS solves this problem by mitigating the definition of the balancing score. Based on the definition, the treatment assignment is conditionally independent of the background variables, CBGPS directly minimize the correlation between the treatment assignment and the covariates after weighting. In specific, the objective of CBGPS is to learn a propensity score based weight so that the weighted correlation between the treatment assignment and the covariates are minimized:

$$\begin{aligned}\mathbb{E}\left(\frac{p(t^*)}{p(t^*|x^*)}t^*x^*\right) &= \int\left\{\int \frac{p(t^*)}{p(t^*|x^*)}t^* dP(t^*|x^*)\right\}x^* dP(x^*) \\ &= \mathbb{E}(t^*)\mathbb{E}(x^*) = 0,\end{aligned} \tag{15}$$

where $p(t^*|x^*)$ is the propensity score, and $\frac{p(t^*)}{p(t^*|x^*)}$ is the balancing weight, t^* and x^* is the treatment assignment and the background variables after centering and orthogonalizing (i.e., normalization). In a nutshell, both CBPS and CBGPS learns the propensity score based sample weight directly towards the covariate balancing goal, which can alleviates negative effect brought by model misspecification of propensity score.

Another drawback of the original IPW estimator is that it might be unstable if the estimated propensity scores are small. If the probability of either treatment assignment is small, the logistic regression model can become unstable around the tails, causing the IPW to also be less stable. To overcome this issue, trimming is routinely employed as a regularization strategy, which eliminates the samples whose propensity scores are less than a pre-defined threshold [85]. However, this approach is highly sensitive to the amount of trimming [94]. Also, theoretical results in [94] show that the small probability of propensity scores and the trimming procedure may result in different non-Gaussian asymptotic distribution of IPW estimator. Based on this observation, a two-way robustness IPW estimation algorithm is proposed in [94]. This method combines subsampling with a local polynomial regression based trimming bias corrector so that it is robust to both small propensity score and the large scale of trimming threshold. An alternative approach to overcome the instability of IPW under small propensity scores is to redesign the sample weight so that the weight is bounded. In [87], the overlap weight is proposed, in which each unit's weight is proportional to the probability of that unit being assigned to the opposite group. In detail, the overlap weight $h(x)$ is defined as $h(x) \propto 1 - e(x)$, where $e(x)$ is the propensity score. The overlap weight is bounded within the interval $[0, 0.5]$, and thus it is less sensitive to the extreme value of

the propensity score. Recent theoretical results show that the overlap weight has the minimum asymptotic variance among all balancing weights [87].

3.1.2 Confounder Balancing. The aforementioned sample re-weighting methods could achieve balance in the sense that the observed variables are considered equally as confounders. However, in real cases, not all the observed variables are confounders. Some of the variables, named as adjustment variables, are only predictive to the outcome, and some others might be irrelevant variables [80]. Adjusting on the adjustment variables by Lasso, although it cannot reduce the bias, helps decrease the variance [20, 132]. While including the irrelevant variables would cause overfitting.

Based on the separateness assumption that the observed variables can be decomposed into confounders, adjusted variables and the irrelevant variables, in [80], the **Data-Driven Variable Decomposition ($\mathbf{D^2VD}$)** algorithm is proposed to distinguish the confounders and adjustment variables, and meanwhile, eliminate the irrelevant variables. In detail, the adjusted outcome is written as:

$$Y^*_{\mathrm{D^2VD}} = \left(Y^F - \phi(\mathbf{z})\right) \frac{W - p(x)}{p(x)(1 - p(x))}, \tag{16}$$

where $\mathbf{z}$ is the adjustment variables. Therefore, the ATE estimator of $\mathrm{D^2VD}$ is:

$$\mathrm{ATE}_{\mathrm{D^2VD}} = \mathbb{E}\left[\left(Y^F - \phi(\mathbf{z})\right) \frac{W - p(x)}{p(x)(1 - p(x))}\right]. \tag{17}$$

To get $\mathrm{ATE}_{\mathrm{D^2VD}}$, $Y^*_{\mathrm{D^2VD}}$ is regressed on all observed variables with parameter α separating the adjustment variables $\mathbf{z}$ from all observed variables and parameter β separating the confounders from all observed variables, i.e., $Y^*_{\mathrm{D^2VD}} = (Y^F - X\alpha) \odot R(\beta)$, where $R(\beta)$ is the the weight and $R(\beta) = \frac{W - e(X)}{e(X)(1-e(X))}$ in which $e(X)$ is parameterized by β. The objective function is l_2 loss between $Y^*_{\mathrm{D^2VD}}$ and ATE value estimated by the linear regression function on all observed variables parameterized by γ, along with sparse regularization to distinguish the confounder, adjusted variables, and irrelevant variables. In detail, the objective function is defined as:

$$\begin{aligned} &\text{minimize } ||(Y^F - X\alpha) \odot R(\beta) - X\gamma||_2^2, \\ &\text{s.t. } \sum_{i=1}^{N} \log(1 + \exp(1 - 2W_i) \cdot X_i\beta)) < \tau, \\ &\quad ||\alpha||_1 \le \lambda, ||\beta||_1 \le \delta, ||\gamma||_1 \le \eta, ||\alpha \odot \beta||_2^2 = 0, \end{aligned} \tag{18}$$

where $R(w)$ is the weight, τ, λ, δ, η is the hyper-parameter. The first condition represents the propensity score estimation error and the next three conditions encourage the sparsity. The last condition, the Hadamard product, ensures the separation of adjusted variables and the confounders.

However, little prior knowledge about the interactions among observed variables is provided in practice, and the data are usually high-dimensional and noisy. To solve this, **Differentiated Confounder Balancing (DCB)** algorithm [79] is proposed to select and differentiate confounders to balance the distributions. Overall, DCB balances the distributions by re-weighting both the samples and confounders.

3.2 Stratification Methods

Stratification, also named as *subclassification* or *blocking* [69], is a representative method to adjust the confounders. The idea of stratification is to adjust the bias that stems from the difference between the treated group and the control group by splitting the entire group into homogeneous subgroups (blocks). Ideally, in each subgroup, the treated group and the control group are similar

under certain measurements over the covariates, therefore, the units in the same subgroup can be viewed as sampled from the data under randomized controlled trials. Based on the homogeneity of each subgroup, the treatment effect within each subgroup (i.e., CATE) can be calculated through the method developed on Randomized Controlled Trials (RCTs) data. After having the CATE of each subgroup, the treatment effect over the interested group can be obtained by combining the CATEs of subgroups belonging to that group, as shown in Equation (8). In the following, we adopt the calculation of ATE as an example. In detail, if we separate the whole dataset into J blocks, the ATE is estimated as:

$$\text{ATE}_{\text{strat}} = \hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j) \right], \tag{19}$$

where $\bar{Y}_t(j)$ and $\bar{Y}_c(j)$ are the average of the treated outcome and control outcome in the j-th block, respectively. $q(j) = \frac{N(j)}{N}$ is the portion of the units in the j-th block to the whole units.

Stratification effectively decreases the bias of ATE estimation compared with the difference-estimator where ATE is estimated as: $\text{ATE}_{\text{diff}} = \hat{\tau}^{diff} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^F - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^F$. In particular, if we assume the outcome is linear with the covariates, i.e., $\mathbb{E}[Y_i(w)|X_i = x] = \alpha + \tau * w + \beta * x$. The bias of the difference-estimator is:

$$\mathbb{E}[\hat{\tau}^{\text{diff}} - \tau | X, W] = (\bar{X}_t - \bar{X}_c)\beta. \tag{20}$$

While, the bias of the stratification estimator is the weighted average of the within-block bias:

$$\mathbb{E}[\hat{\tau}^{\text{strat}} - \tau | X, W] = \left(\sum_{j=1}^{J} q(j) \left(\bar{X}_t(j) - \bar{X}_c(j) \right) \right) \beta. \tag{21}$$

Compared with the difference estimator, the stratification estimator reduces the bias per covariate by the factor:

$$\gamma_k = \frac{\sum_j q(j) \left(\bar{X}_{t,k}(j) - \bar{X}_{c,k}(j) \right)}{\bar{X}_{t,k} - \bar{X}_{c,k}}, \tag{22}$$

where $\bar{X}_{t,k}(j)$ ($\bar{X}_{c,k}(j)$) is the average of k-th covariate of treated (control) group in j-th block, and $\bar{X}_{t,k}$ ($\bar{X}_{c,k}$) is the average of k-th covariate in the whole treated (control) group.

The key component of stratification methods is how to create the blocks and how to combine the created blocks. Equal frequency [122] is a common strategy to create blocks. The equal frequency approach split the block by the appearance probability, such as the propensity score, so that the covariates have the same appearance probability (i.e., the propensity score) in each subgroup (block). The ATE is estimated by the weighted average of each block's CATE, with the weight as the fraction of the units in this block. However, this approach suffers from high variance due to the insufficient overlap between treated and control groups in the blocks whose propensity score is very high or low. To reduce the variance, in [64], the blocks, which are divided according to the propensity score, are re-weighted by the inverse variance of the block-specific treatment effect. Although this method reduces the variance of the equal frequency method, it unavoidably increases the estimation bias.

The stratification methods described above are all splitting the blocks according to the pre-treatment variables. However, in some real-world applications, it is required to compare the outcome conditioned on some post-treatment variables, denoted as S. For example, the "surrogate" markers of disease progression (i.e., intermediate outcome) like CD4 count and measures of viral load in AIDS are the post-treatment variables [48]. In the studies comparing drugs for AIDS patients, the researchers are interested in the effect of AIDS drugs on group with CD4 count lower than 200 cell/mm^3. However, directly comparing the observed outcomes on the group with

$S^{obs} < 200$ is not the true effect because the compared two subgroups: $\{i : W_i = 1, S^{obs} < 200\}$ and $\{j : W_j = 0, S^{obs} < 20\}$, where S^{obs} is the observed post-treatment values, have great discrepency if the treatment has effect on the intermediate results. To solve this, principle stratification [48] constructs the subgroup based on the potential values of the pre-treatment variables. Analogous to the potential outcome defined in Section 2.1, potential pre-treatment variables value, denoted as $S(W = w)$, is the potential value of S under treatment with value w. With the nature assumption that potential value of S is independent of the treatment assignment, the treatment effect of subgroup can be obtained by comparing the outcomes of two sets: $\{Y_i^{obs} : W_i = 1, S_i(W_i = 1) = v_1, S_i(W_i = 0) = v_2\}$ and $\{Y_j^{obs} : W_j = 0, S_j(W_j = 1) = v_1, S_j(W_j = 0) = v_2\}$, where v_1 and v_2 are two post-treatment values. The comparison based on the potential values of post-treatment variables ensures that the compared two set are similar, so that the obtained treatment effect is the true effect.

3.3 Matching Methods

As mentioned previously, missing counterfactuals and confounder bias are two major challenges in treatment effect estimation. Matching based approaches provide a way to estimate the counterfactual and, at the same time, reduce the estimation bias brought by the confounders. In general, the potential outcomes of the i-th unit estimated by matching are [1]:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 1; \end{cases} \quad \hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1; \end{cases} \tag{23}$$

where $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ are the estimated control and treated outcome, $\mathcal{J}(i)$ is the matched neighbors of unit i in the opposite treatment group [13].

The analysis of the matched sample can mimic that of an RCT: one can directly compare outcomes between the treated and control group within the matched sample. In the context of an RCT, one expects that, on average, the distribution of covariates will be similar between treated and control groups. Therefore, matching can be used to reduce or eliminate the effects of confounding when using observational data to estimate treatment effects [13].

3.3.1 Distance Metric. Various distances have been adopted to compare the closeness between units [51], such as the widely used Euclidean distance [126] and Mahalanobis distance [129]. Meanwhile, many matching methods develop their own distance metrics, which can be abstracted as: $D(\mathbf{x}_i, \mathbf{x}_j) = ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2$. The existing distance metrics mainly vary in how to design the transformation function $f(\cdot)$.

Propensity score-based transformation. Original covariates of units can be represented by propensity scores. As a result, the similarity between two units can be directly calculated as: $D(\mathbf{x}_i, \mathbf{x}_j) = |e_i - e_j|$, where e_i, and e_j are the propensity scores of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Later, the linear propensity score based distance metric is also proposed, which is defined as $D(\mathbf{x}_i, \mathbf{x}_j) = |\text{logit}(e_i) - \text{logit}(e_j)|$. This improved version is recommended since it can effectively reduce the bias [151]. Furthermore, the propensity score based distance metric can be combined with other existing distance metrics, which provides a fine-grained comparison. In [129], when the difference of two units' propensity scores is within a certain range, they are further compared with other distances on some key covariates. Under this metric, the closeness of two units contains two criteria: they are relatively close under propensity score measure, and they particularly similar under the comparison of the key covariates [151].

Other transformations. Propensity score only adopts the covariate information, while some other distance metrics are learned by utilizing both the covariates and the outcome information so that the transformed space can preserve more information. One representative metric is the prognosis score [57], which is the estimated control outcome. The transformation function is represented as: $f(x) = \hat{Y}_c$. However, the performance of the prognosis score relies on modeling the relationship between the covariates and control outcomes. Moreover, the prognosis score only takes the control outcome into consideration and ignores the treated outcome. The **Hilbert-Schmidt Independence Criterion based nearest neighbor matching (HSIC-NNM)** proposed in [29] could overcome the drawbacks of prognosis score. HSIC-NNM learns two linear projections for control outcome estimation task and treated outcome estimation task separately. To fully explore the observed control/treated outcome information, the parameters of linear projection is learned by maximizing the nonlinear dependency between the projected subspace and the outcome: $M_w = \arg\max_{M_w} \text{HSIC}(\mathbf{X}_w M_w, Y_w^F) - \mathcal{R}(M_w)$, where $w = 0, 1$ represent the control group and treated group, respectively. $\mathbf{X}_w M_w$ is the transformed subspace with the transformation function as: $f(x) = xM_w$. Y_w^F is the observed control/treated outcome, and $\mathcal{R}$ is the regularization to avoid overfitting. The objective function ensures the learned transformation functions project the original covariates to an information subspace where similar units will have similar outcomes.

Compared with propensity score based distance metric that focuses on balancing, prognosis score and HSIC-NNM focus on embedding the relationship between the transformed space and the observed outcome. These two lines of methods have different advantages, and some recent work tries to integrate these advantages together. In [89], the balanced and nonlinear representation is proposed to project the covariates into a balanced low-dimensional space. In detail, the parameters in the nonlinear transformation function are learned by jointly optimizing the following two objectives: (1) maximizing the differences of noncontiguous-class scatter and within-class scatter so that the units with the same outcome prediction shall have similar representations after transformation; and (2) minimizing the maximum mean discrepancy between the transformed control and outcome group in order to get the balanced space after transformation. A series of works that have similar objectives but vary in balancing regularization have been proposed, such as using the conditional generative adversarial network to ensure the transformation function blocks the treatment assignment information [86, 173].

The methods mentioned above adopt either one or two transformations for treated and control groups separately. Different from the existing method, **Randomized Nearest Neighbor Matching (RNNM)** [90] adopts a number of random linear projections as the transformation function, and the treatment effects are obtained as the median treatment effect by **nearest-neighbor matching (NNM)** in each transformed subspace. The theoretical motivation of this approach is the **Johnson–Lindenstrauss (JL)** lemma, which guarantees that the pairwise similarity information of the points in the high-dimensional space can be preserved through random linear projection. Powered by the JL lemma, RNNM ensembles the treatment effect estimation results of several linear random transformations.

3.3.2 Choosing a Matching Algorithm. After defining the similarity metric, the next step is to find the neighbors. In [26], existing matching algorithms are divided into four essential approaches, including the NNM, caliper, stratification, and kernel, as shown in Figure 3. The most straightforward matching estimator is the NNM. In particular, a unit in the control group is chosen as the matching partner for a treated unit, so that they are closest based on a similarity score (e.g., propensity score). The NNM has several variants like NNM with replacement and NNM without replacement. Treated units are matched to one control, called pair matching or 1-1 matching, or treated units are matched to two controls, called 1-2 matching, and so on. It is a trade-off to

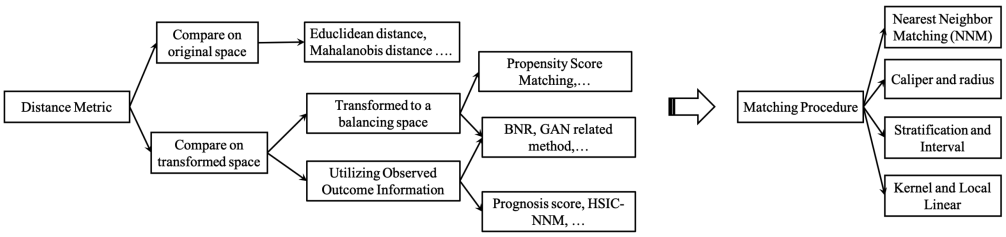

Fig. 3. Categorization of matching methods.

determine the number of neighbors, since a large number of neighbors may result in the treatment effect estimator with high bias but low variance, while small number results in low bias but high variance. It is known, however, that the optimal structure is a full matching in which a treated unit may have one or several controls or one control may have one or several treated units [51].

NNM may have bad matches if the closest partner is far away. One can set a tolerance level on the maximum propensity score distance (caliper) to avoid this problem. Hence, caliper matching is one form of imposing a common support condition.

The stratification matching is to partition the common support of the propensity score into a set of intervals and then to take the mean difference in outcomes between treated and control observations in order to calculate the impact within each interval. This method is also known as interval matching, blocking, and subclassification [124].

The matching algorithms discussed above have in common that only a few observations in the control group are used to create the counterfactual outcome of a treatment observation. Kernel matching and local linear matching are nonparametric matchings that use weighted averages of observations in the control group to create the counterfactual outcome. Thus, one major advantage of these approaches is the lower variance, because we use more information to create a counterfactual outcome.

Here, we also want to introduce another matching method called **Coarsened Exact Matching (CEM)** proposed in [65]. Because either the 1-k matching or the full matching fails to consider the extrapolation region, where few or no reasonable matches exist in the other treatment group, CEM was proposed to handle this problem. CEM first coarsen the selected important covariate, i.e., discretization, and then perform exact matching on the coarsened covariates. For example, if the selected covariates are age (age $>$ 50 is 1, and others are 0) and gender (female as 1, and male as 0). A female patient with age 50 in the treated group is represented by the coarsen covariates as (1, 1). She will only match the patients in the treated group with exactly the same coarsened covariates value. After exact matching, the whole data is separated into two subsets. In one subset, every unit has its exact matched neighbors and it is the opposite in the other subset which contains the units in the extrapolation region. The outcomes of units in the extrapolation region are estimated by the outcome prediction model trained on the matched subset. So far, the treatment effect on the two subsets can be estimated separately, and the final step is to combine the treatment effect on the two subsets by a weighted average.

We have provided several different matching algorithms, but the most important question is how we should select a perfect matching method. Asymptotically all matching methods should yield the same results as the sample size grows and they will become closer to comparing only exact matches [145]. When we only have a small sample size, this choice will be important [60]. There is one trade-off between bias and variance.

3.3.3 Variables to Include. The above two subsections illustrate the key steps in the matching procedure, and in this subsection, we briefly discuss what kinds of variables should be included in the matching, a.k.a feature selection, to improve the matching performance. Many literatures [49, 60, 128] suggest to include as many variables that are related to the treatment assignment and the outcome as possible, in order to satisfy the strong ignorablity assumption. However, post-treatment variables, which are the variables affected by the treatment assignment, should be excluded in the matching procedure [123]. Moreover, besides the post-treatment variables, researchers also suggest excluding the instrumental variables [106, 170], because they tend to amplify the bias of treatment effect estimator.

3.4 Tree-based Methods

Another popular method in causal inference is based on decision tree learning, which is one of the predictive modeling approaches. The decision tree is a nonparametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data.

Tree models where the target variable is discrete are called classification trees with prediction error measured based on misclassification cost. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable is continuous are called regression trees with prediction error measured by the squared difference between the observed and predicted values. The term **Classification And Regression Tree (CART)** analysis is an umbrella term used to refer to both of the above procedures [24]. In CART model, the data space is partitioned and a simple prediction model for each partitioned space is fitted, and therefore every partitioning can be represented graphically as a decision tree [92].

For estimating heterogeneity in causal effects, a data-driven approach [11] based on CART is provided to partition the data into subpopulations that differ in the magnitude of their treatment effects. The valid confidence intervals can be created for treatment effects, even with many covariates relative to the sample size, and without "sparsity" assumptions. This approach is different from conventional CART in two aspects. First, it focuses on estimating conditional average treatment effects instead of directly predicting outcomes as in the conventional CART. Second, different samples are used for constructing the partition and estimating the effects of each subpopulation, which is referred to as an honest estimation. However, in the conventional CART, the same samples are used for these two tasks.

In CART, a tree is built up until a splitting tolerance is reached. There is only one tree, and it is grown and pruned as needed. However, BART is an ensemble of trees, so it is more comparable to random forests. A Bayesian "sum-of-trees" model called **Bayesian Additive Regression Trees (BART)** is developed in [33] [34]. Every tree in BART model is a weak learner, and it is constrained by a regularization prior. Information can be extracted from the posterior by a Bayesian backfitting MCMC algorithm. BART is a nonparametric Bayesian regression model, which uses dimensionally adaptive random basis elements. Let W be a binary tree which has a set of interior node decision rules and terminal nodes, and let $M = \{\mu_1, \mu_2, ..., \mu_B\}$ be parameters associated with each of the B terminal nodes for W. We use $g(x; W, M)$ to assign a $\mu_b \in M$ to input vector x. The sum-of-trees model can be expressed as:

$$Y = g(x; W_1, M_1) + g(x; W_2, M_2) + \cdots + g(x; W_m, M_m) + \varepsilon, \tag{24}$$

$$\varepsilon \sim N(0, \sigma^2). \tag{25}$$

BART has a couple of advantages. It is very easy to implement and only needs to plug in the outcome, treatment assignment, and confounding covariates. In addition, it does not require any

information about how these variables are parametrically related so that it requires less guess when fitting the model. Moreover, it can deal with a mass of predictors, yield coherent uncertainty intervals, and handle continuous treatment variables and missing data [61].

BART is proposed to estimate the average causal effects. In fact, it can also be used to estimate individual-level causal effects. BART not only can easily identify the heterogeneous treatment effects, but also get more accurate estimates of average treatment effects compared to other methods like propensity score matching, propensity score weighting, and regression adjustment in the nonlinear simulation situations examined [61].

In most previous methods, the prior distribution over treatment effects is always induced indirectly, which is difficult to be attained. A flexible sum of regression trees (i.e., a forest) can address this issue by modeling a response variable as a function of a binary treatment indicator and a vector of control variables [56]. This approach interpolates between two extremes: entirely and separately modeling the conditional means of treatment and control, or only the treating treatment assignment as another covariate.

Random forest is a classifier consisting of a combination of tree predictors, in which each tree depends on a random vector that is independently sampled and has the identical distribution for all trees [23]. This model can also be extended to estimate heterogeneous treatment effects based on Breiman's random forest algorithm [162]. Trees and forests can be considered as nearest neighbor methods with an adaptive neighborhood metric. Tree-based methods seek to find training examples that are close to a point x, but now closeness is defined with respect to a decision tree. And the closest points to x are those that fall in the same leaf as it. The advantage of using trees is that their leaves can be narrower along with the directions where the signal is changing fast and wider along with the other directions, potentially leading to a substantial increase in power when the dimension of the feature space is even moderately large.

The tree-based framework also can be extended to uni- or multi-dimensional treatments [163]. Each dimension can be discrete or continuous. A tree structure is used to specify the relationship between user characteristics and the corresponding treatment. This tree-based framework is robust to model misspecification and highly flexible with minimal manual tuning.

3.5 Representation Learning Methods

Representation learning is learning the representations of input data typically by transforming the original covariates or extracting features from the covariates space. Focusing specifically on deep learning, the composition of multiple nonlinear transformations can yield more abstract and ultimately more useful representations [17]. Compared with traditional machine learning approaches in causal inference, deep representation learning models are capable of automatically searching for features that are correlated and combine them together to enable more effective and more accurate counterfactual estimation while in the traditional machine learning approach, features need to be identified accurately by users. Meanwhile, there also exists some challenges that need to be addressed in deep representation learning. For example, the amount of data needed for deep representation learning is much higher than other machine learning methods; the "Black Boxes" deep structure is less interpretable and it is very difficult to look inside of it to understand how it works; overfitting problem always happens when an algorithm utilizes the deep structure to learn the details and noise so well in the training data that it negatively impacts the performance of the model in the whole population. Thus far, significant advances in deep representation learning based methods have been made to overcome the challenges in causal effect estimation with observational data. We categorize the deep representation learning based methods into domain adaptation based, matching based, and continual learning based methods.

3.5.1 Domain Adaptation Based on Representation Learning. The most basic assumption used in statistical learning theory is that training data and test data are drawn from the same distribution. However, in most practical cases, the test data are drawn from a distribution that is only related, but not identical, to the distribution of the training data. In causal inference, this is also a big challenge. Unlike the randomized control trials, the mechanism of treatment assignment is not explicit in observational data. Therefore, interventions of interest are not independent of the property of the subjects. For example, in an observational study of the treatment effect of a medicine, the medicine is assigned to individuals based on several factors, including the known confounders and some unknown confounders. As a result, the counterfactual distribution will generally be different from the factual distribution. Thus, it is necessary to predict counterfactual outcomes by learning from the factual data, which converts the causal inference problem to a domain adaptation problem.

Extracting effective feature representations is critical for domain adaptation. A model [16] with a generalization bound is proposed to formalize this intuition theoretically, which can not only explicitly minimize the difference between the source and target domains, but also maximize the margin of the training set. Building on this work [16], the discrepancy distance between distributions is tailored to adaptation problems with arbitrary loss functions [95]. In the following discussions, the discrepancy distance plays an important role in addressing the domain adaptation problem in causal inference.

So far, we can see a clear connection between counterfactual inference and domain adaptation. An intuitive idea is to enforce the similarity between the distributions of different treatment groups in the representation space. The learned representations trade-off three objectives: (1) low-error prediction over the factual representation; (2) low-error prediction over counterfactual outcomes by taking into account relevant factual outcomes; and (3) the distance between the distribution of treatment population and that of control population [70]. Following this motivation, [139] gives a simple and intuitive generalization-error bound. It shows that the expected ITE estimation error of representation is bounded by a sum of the standard generalization-error of that representation and the distance between the treated and control distributions based on representation. **Integral probability metric (IPM)** is used to measure the distances between distributions, and explicit bounds are derived for the Wasserstein distance and Maximum Mean Discrepancy distance. The goal is to find a representation $\Phi : X \to R$ and hypothesis $h : X \times \{0, 1\} \to Y$ that minimizes the following objective function:

$$\min_{h,\Phi} \frac{1}{n} \sum_{i=1}^{n} r_i \cdot L(h(\Phi(x_i), W_i), y_i) + \lambda \cdot R(h) + \alpha \cdot IPM_G(\{\Phi(x_i)\})_{i:W_i=0}, \{\Phi(x_i)\})_{i:W_i=1}), \tag{26}$$

where $r_i = \frac{W_i}{2u} + \frac{1-W_i}{2(1-u)}$, $u = \frac{1}{n}\sum_{i=1}^{n} W_i$, and the weight r_i compensates for the difference in treatment group size. R is a model complexity term. Given two probability density functions p, q defined over $S \subseteq R^d$, and a function family G of functions $g : S \to R$, the IPM is defined as:

$$IPM_G(p, q) := \sup_{g \in G} | \int_S g(s)(p(s) - q(s))ds|. \tag{27}$$

This model allows for learning complex nonlinear representations and hypotheses with large flexibility. When the dimension of Φ is high, it risks losing the influence of t on h if the concatenation of Φ and W is treated as input. To address this problem, one approach is to parameterize $h_1(\Phi)$ and $h_0(\Phi)$ as two separate "heads" of the joint network. $h_1(\Phi)$ is used to estimate the outcome under treatment and $h_0(\Phi)$ is for the control group. Each sample is used to update only the head corresponding to the observed treatment. The advantage is that statistical power is shared in the

common representation layers and the influence of treatment is retained in the separate heads [139]. This model can also be extended to any number of treatments, as described in the perfect match approach [136]. Following this idea, a few improved models have been proposed and discussed. For example, [71] brings together shift-invariant representation learning and re-weighting methods. [59] presents a new context-aware weighting scheme based on the importance sampling technique, on top of representation learning, to alleviate the selection bias problem in ITE estimation.

Existing ITE estimation methods mainly focus on balancing the distributions of control and treated groups, but ignore the local similarity information that provides meaningful constraints on the ITE estimation. In [171, 172], a local **similarity preserved individual treatment effect (SITE)** estimation method is proposed based on deep representation learning. SITE preserves local similarity and balances data distributions simultaneously. The framework of SITE contains five major components: representation network, triplet pairs selection, **position-dependent deep metric (PDDM)**, **middle point distance minimization (MPDM)**, and the outcome prediction network. To improve the model efficiency, SITE takes input units in a mini-batch fashion, and triplet pairs could be selected from every mini-batch. The representation network learns latent embeddings for the input units. With the selected triplet pairs, PDDM and MPDM can preserve the local similarity information and meanwhile achieve the balanced distributions in the latent space. Finally, the embeddings of mini-batch are fed forward to a dichotomous outcome prediction network to get the potential outcomes. The loss function of SITE is as follows:

$$L = L_{FL} + \beta L_{PDDM} + \gamma L_{MPDM} + \lambda ||M||_2, \tag{28}$$

where L_{FL} is the factual loss between the estimated and observed factual outcomes. L_{PDDM} and L_{MPDM} are the loss functions for PDDM and MPDM, respectively. The last term is L_2 regularization on model parameters M (except the bias term).

Most models focus on covariates with numerical values, while how to handle covariates with textual information for treatment effect estimation is still an open question. One major challenge is how to filter out the nearly instrumental variables which are the variables more predictive to the treatment than the outcome. Conditioning on those variables to estimate the treatment effect would amplify the estimation bias. To address this challenge, a **conditional treatment-adversarial learning based matching (CTAM)** method is proposed in [173]. CTAM incorporates the treatment-adversarial learning to filter out the information related to nearly instrumental variables when learning the representations, and then it performs matching among the learned representations to estimate the treatment effects. The CTAM contains three major components: text processing, representation learning, and conditional treatment discriminator. Through the text processing component, the original text is transformed into vectorized representation S. After that, S is concatenated with the non-textual covariates X to construct a unified feature vector, which is then fed into the representation neural network to get the latent representation Z. After learning the representation, Z, together with potential outcomes Y, are fed into the conditional treatment discriminator. During the training procedures, the representation learner plays a minimax game with the conditional treatment discriminator: By preventing the discriminator from assigning correct treatment, the representation learner can filter out the information related to nearly instrumental variables. The final matching procedure is performed in the representation space Z. The conditional treatment-adversarial learning helps reduce the bias of treatment effect estimation.

3.5.2 Matching Based on Representation Learning. Compared to the above regression-based methods after representation learning, matching methods based on representation learning are

more interpretable, because any sample's counterfactual outcome is directly set to be the factual outcome of its nearest neighbor in the group receiving the opposite treatment. NNM sets the counterfactual outcome of any treatment (control) sample to be equal to the factual outcome of its nearest neighbor in the control (treatment) group. Although being simple, flexible, and interpretable, most NNM approaches could be easily misled by variables that do not affect the outcome. To address this challenge, matching can be performed on subspaces that are predictive of the outcome variable for both the treatment group and the control group. Applying NNM in the learned subspaces leads to a more accurate estimation of the counterfactual outcomes and therefore the accurate estimation of treatment effects. For example, one work [29] estimates the counterfactual outcomes of treatment samples by learning a projection matrix that maximizes the nonlinear dependence between the subspace and outcome variable for control samples. Then it directly applies the learned projection matrix to all the samples and finds every treatment sample's matched control sample in the subspace. In addition, another work [35] performs matching in the selective and balanced representation space to estimate treatment effects. It seamlessly integrates deep feature selection and deep representation learning for causal inference together. In the feature selection and representation learning, that one-to-one feature selection layer at the input level selects which variables are input into the neural network, which makes the deep neutral network more interpretable.

3.5.3 Continual Learning Based on Representation Learning. Although significant advances have been made to overcome the challenges in causal effect estimation with observational data, the existing representation learning methods only focus on the source-specific and stationary observational data. Such learning strategies assume that all observational data are already available during the training phase and from the only one source. This assumption is unsubstantial in practice due to two reasons. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. For instance, the number of electronic medical records in one hospital is growing every day, or the electronic medical records for one disease may be from different hospitals or even different countries. This characteristic implies that one cannot have access to all observational data at one time point and from one single source. The second reason is based on the realistic consideration of accessibility. For example, when the new observational are available, if we want to refine the model previously trained by original data, maybe the original training data are no longer accessible due to a variety of reasons, e.g., lost, proprietary, too large to store, or privacy constraint. This practical concern of accessibility is ubiquitous in various academic and industrial applications. A continual causal effect representation learning method [8] is proposed for estimating causal effect with observational data, which are incrementally available from non-stationary data distributions. Instead of having access to all seen observational data, it incorporates feature representation distillation to preserve the knowledge learned from previous observational data. Besides, aiming at solving the selection bias between treatment and control groups, it adopts one representation transformation function, which maps partial original feature representations into new feature representation space and makes the global feature representation space balanced with respect to treatment and control groups.

3.6 Multi-task Learning Methods

The treatment group and control group always share some common features except for their idiosyncratic characteristics. Naturally, causal inference can be conceptualized as a multi-task learning problem with a set of shared layers for the treated group and control group together, and a set of specific layers for the treated group and control group separately. The impact of selection bias in the multi-task learning problem can be alleviated via a propensity-dropout regularization

scheme [5], in which the network is thinned for every training example via a dropout probability that depends on the associated propensity score. The dropout probability is higher for subjects with features that belong in a region of poor overlap in the feature space between the treatment and control group.

The Bayesian method also can be extended under the multi-task model. A nonparametric Bayesian method [4] uses a multi-task Gaussian process with a linear coregionalization kernel as a prior over the vector-valued reproducing kernel Hilbert space. The Bayesian approach allows computing individualized measures of confidence in our estimates via pointwise credible intervals, which are crucial for realizing the full potential of precision medicine. The impact of selection bias is alleviated via a risk-based empirical Bayes method for adapting the multi-task GP prior, which jointly minimizes the empirical error in factual outcomes and the uncertainty in counterfactual outcomes.

The multi-task model can be extended to multiple treatments even with continuous parameters in each treatment. The **dose response network (DRNet)** architecture [135] with shared base layers, N_W intermediary treatment layers, and $N_W \times E$ heads for the multiple treatment setting with an associated dosage parameter s. The shared base layers are trained on all samples, and the treatment layers are only trained on samples from their respective treatment category. Each treatment layer is further subdivided into E head layers. Each head layer is assigned a dosage stratum that subdivides the range of potential dosages $[a_t, b_t]$ into E partitions of equal width $\frac{b-a}{E}$.

3.7 Meta-Learning Methods

When designing the heterogeneous treatment effect estimation algorithms, two key factors should be considered: (1) control the confounders, i.e., eliminate the spurious correlation between the confounder and the outcome; and (2) give an accurate expression of the CATE estimation [99]. The methods mentioned in the previous sections seek to satisfy the two requirements simultaneously, while meta-learning based algorithms separate them into two steps. In general, the meta-learning based algorithms have the following procedures: (1) estimate the conditional mean outcome $\mathbb{E}[Y|X = x]$, and the prediction model learned in this step is the base learner. (2) Derive the CATE estimator based on the difference of results obtained from step (1). Existing meta-learning methods include T-learner [81], S-learner [81], X-learner [81], U-learner [99], and R-learner [99], which are introduced in the following.

In detail, the T-learner [81] adopts two trees to estimate the conditional treated/control outcomes, which are denoted as $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ and $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$, respectively. Let $\hat{\mu_0}(x)$ and $\hat{\mu_0}(x)$ denote the trained tree model on the control/treated group. Then the CATE of T-learner estimation is obtained as: $\hat{\tau}_T(x) = \hat{\mu_1}(x) - \hat{\mu_0}(x)$. T-learner trains two base models for control and treated groups (the name "T" comes from two base model), while S-learner[81] views the treatment assignment as one feature and estimate the combined outcome as: $\mu(x, w) = \mathbb{E}[Y^F|X = x, W = w]$ (The name "S" denotes single). $\mu(x, w)$ can be any base model, and we denote the trained model as $\hat{\mu}(x, w)$. The CATE estimator provided by S-learner is then given as: $\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$.

However, T-learner and S-learner highly rely on the performance of the trained base models. When the number of units in two groups are extremely unbalanced (i.e., the number of one group is much larger than the other), the performance of the base model trained on the small group would be poor. To overcome this problem, X-learner [81] is proposed, which adopts information from the control group to give a better estimator on the treated group and vice versa. The cross-group information usage is where X-learner comes from, and the X denotes "cross group." In detail, X-learner contains three key steps. The first step of X-learner is the same as T-learner, and the

trained base learners are denoted as $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$. In the second step, X-learner calculates the difference between the observed outcome and the estimated outcome as the imputed treatment effect: In the control group, the difference is the estimated treated outcome subtracts the observed control outcome, denoted as $\hat{D}_i^C = \hat{\mu}_1(x) - Y^F$; Similarly, in the treated group, the difference is formulated as $\hat{D}_i^T = Y^F - \hat{\mu}_0(x)$. After the difference calculation, the dataset is transformed into two groups with imputed treatment effect: control group: $(X_C, \hat{D}^C)$ and treated group: $(X_T, \hat{D}^T)$. On two imputed datasets, the two base learners of treatment effect $\tau_1(x)(\tau_0(x))$ are trained with $X_C(X_T)$ as the input and $\hat{D}^C(\hat{D}^T)$ as the output. The last step is to combine the two CATE estimators by weighted average: $\tau_X(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$, where $g(x)$ is the weighting function ranging from 0 to 1. Overall, with the cross information usage and the weighted combination of two CATE base estimator, X-learners can handle the case where the number of units in two groups are unbalanced [81].

Different from the regular loss function adopted in X-learner, R-learner, proposed in [99] designs the loss function for CATE estimator based on the Robinson transformation [120]. The character "R" in R-learner denotes the Robinson transformation. The Robinson transformation can be derived by rewriting the observed outcome and the conditional outcome: Rewrite the observed outcome as:

$$Y_i(W = w_i) = \hat{\mu}_0(x_i) + w_i * \tau(x_i) + \epsilon_i(w_i), \tag{29}$$

where $\hat{\mu}_0$ is the already-trained control outcome estimator(base learner), $\tau(x_i)$ is the CATE estimator, and $E[\epsilon_i(w_i)|x_i, w_i] = 0$ (under ignorability). The conditional mean outcome can be also rewritten as:

$$\hat{m}(x_i) = E[Y|X] = \hat{\mu}_0(x_i) + \hat{e}(x_i) * \tau(x_i), \tag{30}$$

where $\hat{e}(x)$ is the already-trained propensity score estimator(base learner). Robinson transformation is obtained by subtracting Equations (29) and (30):

$$Y_i^F - \hat{m}(x_i) = (w_i - \hat{e}(x_i))\tau(x_i) + \epsilon(w_i). \tag{31}$$

Based on the Robinson transformation, a good CATE estimator should minimize the difference between $Y_i^F - \hat{m}(x_i)$ and $(w_i - \hat{e}(x_i))\tau(x_i)$. Therefore, the objective function of R-learner is as follows:

$$\tau(\cdot) = \arg\min_\tau \left\{ \frac{1}{n} \sum_{i=1}^{n} \left((Y_i^F - \hat{m}(x_i)) - (w_i - \hat{e}(x_i))\, \tau(x_i) \right)^2 + \Lambda(\tau(\cdot)) \right\}, \tag{32}$$

where $\hat{m}(x_i)$ and $\hat{e}(x_i)$ are pre-trained outcome estimator and propensity score estimator, and $\Lambda(\tau(\cdot))$ is the regularization on $\tau(\cdot)$.

4 METHODS RELAXING THREE ASSUMPTIONS

In Section 3, the causal inference methods based on three assumptions have been introduced in detail, which are theSUTVA, ignorability assumption, and positivity assumption. However, in practice, for some specific applications like social media analysis, which involves dependent network information, special data types (e.g., time series data), or particular conditions (e.g., the existence of unobserved confounders), these three assumptions cannot always hold. In this section, the methods that try to relax certain assumptions will be discussed.

4.1 Relaxing Stable Unit Treatment Value Assumption

SUTVA states that the potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. This assumption mainly focuses on two aspects: (1)

units are independent and identically distributed (i.i.d.); and (2) there only exists a single level for each treatment. An extensive literature exists on making causal inferences under SUTVA, but when considering many real-world situations, it may not always be the case. In the following, SUTVA will be discussed from these two aspects.

The assumption of independent and identically distributed samples is ubiquitous in most causal inference methods, but this assumption cannot hold in many research areas, such as social media analytics [54] [140], herd immunity, and signal processing [153, 161]]. Causal inference in non-i.i.d. contexts is challenging due to the presence of both unobserved confounding and data dependence. For example, in social networks, subjects are connected and influenced by each other.

For such network data, SUTVA cannot hold anymore. Under this situation, instances are inherently interconnected with each other through the network structure and hence their features are not independent identically distributed samples drawn from a certain distribution. Applying Graph Convolutional Networks into a causal inference model is an approach to handle the network data [54]. In particular, the original features of subjects and the network structure are mapped to a representation space to get the representation of confounders. Furthermore, the potential outcomes could be inferred using treatment assignments and confounder representations.

The dependence on data often leads to interference because some subjects' treatments can affect others' outcomes [63, 101]. This difficulty can impede the identification of causal parameters of interest. Extensive work has been developed on the identification and estimation of causal parameters under interference [63, 101, 108, 157]. For this problem, a strategy proposed by [142] is to use segregated graphs [144], a generalization of latent projection mixed graphs [160], to represent causal models.

Modeling time series data is another important problem in causal inference, which does not satisfy the independent and identically distributed assumption. Most of the existing methods use regression models for this problem but the accuracy of inference depends greatly on whether the model fits the data. Therefore, selecting a right and appropriate regression model is of crucial importance, but in practice, it is not easy to find the perfect one. [32] proposes a supervised learning framework that uses a classifier to replace regression models. It presents a feature representation that employs the distance between the conditional distributions given past variable values and shows experimentally that the feature representation provides sufficiently different feature vectors for time series with different causal relationships. For the time series data, another issue that needs to be considered is hidden confounders. A time series deconfounder [18] was developed, which leverages the assignment of multiple treatments over time to enable the estimation of treatment effects even in the presence of hidden confounders. This time series deconfounder uses a recurrent neural network architecture with multi-task output to build a factor model over time and infer substitute confounders, which render the assigned treatments conditionally independent. Then it performs causal inference using the substitute confounders.

For the second direction in SUTVA assumption, it assumes that there only exists one version for each treatment. However, if adding one continuous parameter into the treatment, this assumption cannot hold anymore. For example, estimating individual dose-response curves for a couple of treatments requires adding an associated dosage parameter (categorical or continuous) for each treatment. Under this situation, for each treatment, it will have multiple versions for categorical dosage parameters or infinite versions for continuous dosage parameters. One way to solve this problem is to convert the continuous dosage into a categorical variable and then treat every medication with a specific dosage as one new treatment, so that it will satisfy the SUTVA assumption again [135].

Another example that breaks the SUTVA is the dynamic treatment regime, which consists of a sequence of decision rules, one per stage of intervention [27]. One useful application of dynamic treatment is precision medication. It includes more individualization to adjust which type of treatment should be used, or how many the dosage is best in response to the patient's background characteristics, the illness severity, and other heterogeneity, aiming to get the optimal treatment strategy. These heterogeneities are called tailoring variables. To get a useful dynamic treatment regime, [84] introduces one "biased coin adaptive within-subject" design. Then, [97] presents one general framework of this type of design, which uses sequential multiple assignment randomized trials for developing decision rules in that each individual may be randomized multiple times and the multiple randomizations occur sequentially over time.

For estimating optimal dynamic decision rules from observational data, Q [166, 167] and A [96, 118] learning are two main approaches for estimating the optimal dynamic treatment regime. Q in Q-learning denotes "quality." Q-learning is a model-free reinforcement learning algorithm, which employs posited regression models for estimating outcome at each decision point given units' information. In advantage learning (A-learning), models are posited only for the part of the regression including contrasts among treatments and for the probability of observed treatment assignment at each decision point, given units' information. Both methods are implemented through a backward recursive fitting procedure that is related to dynamic programming [15].

4.2 Relaxing Unconfoundedness Assumption

The ignorability assumption is also named as the unconfoundedness assumption. Given the background variable, X, the treatment assignment W is independent to the potential outcomes, i.e., $W \perp\!\!\!\perp Y(W = 0), Y(W = 1)|X$. With this unconfoundedness assumption, for the units with the same background variable X, their treatment assignment can be viewed as random. Obviously, identifying and collecting all of the background variables is impossible, and this assumption is very difficult to satisfy. For example, in an observational study that tries to estimate the individual treatment effect of a medicine, instead of randomized experiments, the medicine is assigned to individuals based on a series of factors. Some factors (e.g., socioeconomic status) are challenging to measure and therefore become hidden confounders. Existing work overwhelmingly relies on the unconfoundedness assumption that all confounders can be measured. However, this assumption might be untenable in practice. In the above example, units' demographic attributes, such as their home address, consumption ability, or employment status, may be the proxies for socioeconomic status. Leveraging big data, it is possible to find a proxy for the latent and unobserved confounders.

Variational autoencoder has been used to infer the complex non-linear relationships between the observed confounders and joint distribution of the latent confounders, treatment assignment, and outcomes [93]. The joint distribution of the latent confounders and the observed confounders can be approximately recovered from the observations. An alternative way is to capture their patterns and control their influence by incorporating the underlying network information. Network information is also a reasonable proxy for the unobserved confounding. [54] applies GCN on network information to get the representation of hidden confounders. Moreover, in [53], graph attention layers are used to map the observed features in networked observational data to the D-dimensional space of partial latent confounders, by capturing the unknown edge weights in the real-world networked observational data.

An interesting insight mentioned in [159] is that, even if the confounders are observed, it does not mean all the information they contain is useful to infer the causal effect. Instead, requiring the part of confounders actually used by the estimator is sufficient. Therefore, if a good predictive model for the treatment can be built, one may only need to plug the outputs into a causal

effect estimate directly, without any need to learn the whole true confounders. In [159], the main idea is to reduce the causal estimation problem to a semi-supervised prediction of both the treatments and outcomes. Networks admit high-quality embedding models that can be used for this semi-supervised prediction. In addition, embedding methods can also offer an alternative to fully specified generative models.

Only using observational data to solve the confoundings problem is always difficult. The alternative way is to combine the experimental data and observational data together. In [73], limited experimental data is used to correct the hidden confounding in causal effect models trained on larger observational data, even if the observational data do not fully overlap with the experimental data. This method makes strictly weaker assumptions than existing approaches.

For estimating treatment effects from longitudinal observational data, existing methods usually assume that there are no hidden confounders. This assumption is not testable in practice and, if it does not hold, leads to biased estimates. [18] infers substitute confounders that render the assigned treatments conditionally independent. Then it performs causal inference using the substitute confounders. This method can help estimate treatment effects for time series data in the presence of hidden confounders.

The above methods all aim to solve the problems about the observed and unobserved confounders. Are there any other ways to get around the unconfoundedness assumption and conduct causal inference? One way is to use instrumental variables that only affect the treatment assignment but not the outcome variable. Changes in the instrumental variables would lead to a different assignment of treatment. [58] broke instrumental variables analysis into two supervised stages that can each be targeted with deep networks. It models the conditional distribution of the treatment variable given the instruments and covariates, and then employs a loss function involving integration over the conditional treatment distribution. The deep instrumental variables framework also takes advantage of existing supervised learning techniques to estimate causal effects.

4.3 Relaxing Positivity Assumption

The positivity assumption, also known as covariate overlap or common support, is a necessary assumption for the identification of treatment effect in the observational study. However, little literature discusses the satisfaction of this assumption in the high dimensional datasets. [38] argues that the positivity assumption is a strong assumption and is more difficult to be satisfied in the high-dimensional datasets. To support the claim, the implication of the strict overlap assumption is explored, and it shows that strict overlap restricts the general discrepancies between the control and treated covariates. Therefore, the positivity assumption is stronger than the investigator expected. Based on the above implication, methods that eliminate the information about the treatment assignment while still hold the unconfounderness assumption are recommended, such as trimming [36, 110, 122] that drops the records in the region without overlap, and instrumental variable adjustment methods [41, 98, 106] that eliminate the instrumental variables from covariates.

5 GUIDELINE ABOUT EXPERIMENT

In this section, we provide the related experimental information, including the available datasets that are commonly adopted in the experiments, and the open-source codes of the methods mentioned in the previous two sections.

5.1 Available Datasets

Because the counterfactual outcome can never be observed, it is hard to find the dataset that perfectly satisfies the requirements of the experiment that it is an observational dataset with the

ground truth ATE (or ITE) available. The datasets used in the literature are often semi-synthetic datasets. Some datasets, such as the IHDP dataset, are obtained from the randomized dataset by generating their observed outcome according to a certain generation process and removing a biased subset to mimic the selection bias in the observational dataset. Some datasets, such as the Jobs dataset, combine the randomized dataset and the observational control dataset together to create the selection bias. The details and download links of the available benchmark datasets are provided in Table 2.

IHDP. This dataset is a commonly adopted benchmark dataset. This dataset is generated based on the randomized controlled experiment conducted by Infant Health and Development Program [25], whose targets are low-birthweight and premature infants. The pre-treatment covariates measure the aspects of the children and their mothers, such as birth weight, head circumference, neonatal health index, prenatal care, mother's age, education, drugs, and alcohol. In the treated group, the infants are provided with both intensive high-quality childcare and specialist home visits [61]. The outcome is the infants' cognitive test score and can be simulated through the NPCI package.[1] Besides, a biased subset of the treated group is required to be removed to simulate the selection bias. An example of the IHDP dataset whose outcome is simulated by the setting "A" of NPCI package can be downloaded from http://www.mit.edu/~fredrikj/files/ihdp_100.tar.gz.

Jobs. The jobs datasets used in the observational study [39, 40, 139] is the combination of Lalonde experiment data and the PSID comparison group. Both Lalonde and PSID datasets can be downloaded from NBER website.[2] The pre-treatment covariates are eight variables such as age, education, ethnicity, as well as earnings in 1974 and 1975. The people in the treated group take part in the job training while in the control group are not. The outcome is employment status.

Twins. Twins dataset is first introduced in [93] and is adopt by various observational studies [93, 171, 174]. Twins dataset is constructed on the data of twins birth in the USA between 1989 and 1991 [6].[3] In [93], the twins whose gender is the same and weight is less than 2000g are selected into records. For each twin pair, the pre-treatment covariates measure the pregnancy, twins birth, and the parents, such as the number of gestation weeks before birth, the quality of care during pregnancy, pregnancy risk factors (anemia, alcohol use, tobacco use, etc.), adequacy of care, and residence. The outcome is 1-year mortality. The treatment is being the heavier one in the twins, and the outcome is 1-year mortality. In the twins dataset, both treated (the heavier one in the twin) and control (the lighter one in the twin) outcomes are observed. The treatment assignment is usually defined by the users to simulate the selection bias. For example, in [93, 174], the selection bias is created by the following procedure: $W_i|\mathbf{X}_i \sim Bern(Sigmoid(\mathbf{w}^{'}\mathbf{X}_i) + n)$, where $\mathbf{w} \sim U(-0.1, 0.1)^{40\times 1}$ and $n \sim \mathbf{N}(0, 0.1)$.

ACIC datasets. Starting from 2016, every year, the Atlantic Causal Inference Conference holds the causal inference data analysis challenge, which provides some datasets targeting different causal inference problems.

2016 Challenge: The goal of ACIC 2016 Challenge is to better understand which approaches to causal inference perform well in particular observational study settings.[4] The datasets contain 77 datasets with varying degrees of non-linearity, sparsity, correlation between treatment assignment and outcome, non-linearity of treatment effect, and overlapping. The covariates are real-world data from the Infant Health and Development Program dataset [25], which consists of 58 variables. The

[1] https://github.com/vdorie/npci.
[2] http://users.nber.org/~rdehejia/data/nswdata2.html.
[3] www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html.
[4] https://jenniferhill7.wixsite.com/acic-2016/competition.

Table 2. Summary of Available Datasets for Causal Inference

Dataset	Source	Data description	Link
IHDP	It is generated based on the randomized controlled experiment conducted by Infant Health and Development Program [25].	747 units (139 treated, 608 control), 25 pre-treatment variables and binary treatment	http://www.mit.edu/~fredrikj/files/ihdp_100.tar.gz
Jobs	It is the combination of Lalonde experiment data and the PSID comparison group [39, 40, 139].	LaLonde experimental 722 units (297 treated, 425 control), PSID comparison group (2490 control). 8 pre-treatment and binary treatment	http://users.nber.org/~rdehejia/data/nswdata2.html
Twins	Twins dataset is constructed on the data of twins birth in the USA between 1989-1991 [6].	11,984 pairs of twins, 46 pre-treatment and binary treatment	www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html
ACIC 2016	The covariates are real-world data from the Infant Health and Development Program dataset [25].	77 datasets, 58 variables and binary treatments	https://jenniferhill7.wixsite.com/acic-2016/competition
ACIC 2017	The covariates are from Infant Health and Development Program dataset [25].	8,000 datasets 8 variables	https://www.niss.org/events/2017-atlantic-causal-inference-conference-acic
ACIC 2018	Linked Births and Infant Deaths Database (LBIDD)	24 simulations with sizes from 1000 to 50000 and binary treatments	https://www.synapse.org/ACIC2018Challenge
ACIC 2019	There are 3200 low dimensional datasets, and 3200 high-dimensional datasets.	Several datasets with different variables size and record size, and the R code for data generation is available.	https://sites.google.com/view/acic2019datachallenge/data-challenge?authuser=0
IBM causal inference data	This dataset is created in [143]. The framework includes unlabeled data for prediction, labeled data for validation, and code for evaluation of algorithm.	This dataset uses the cohort of 100K samples in Linked Births and Infant Deaths Database (LBIDD) as the fundamental set of covariates.	https://github.com/IBM-HRL-MLHLS/IBM-Causal-Inference-Benchmarking-Framework
BlogCatalog	BlogCatalog is an online community where users post blogs. BlogCatalog dataset used in [54].	There are 5,196 users, 171,743 social relations, 8,189 features, and 6 categories of blogs.	http://networkrepository.com/soc-BlogCatalog.php
Flickr	Flickr is an image sharing website where users interact with each other via photo sharing under predefined categories [54].	There are 7,575 users, 12,047 features, 239,738 social relations, and 9 classes.	https://snap.stanford.edu/data/web-flickr.html
News	The News benchmark includes randomly sampled news articles from the NY Times corpus.	5000 samples, Multiple treatments (2, 4, 8, and 16)	https://polybox.ethz.ch/index.php/s/fRQZREF528AfD5Z
MVICU	The data is sourced from the publicly available MIMIC III database [130].	3 treatments	https://mimic.physionet.org/
TCGA	This dataset is generated by The Cancer Genome Atlas Pan-Cancer analysis project. [168].	Gene expression from different cancers in 9,659 individuals. 3 available clinical treatments	https://polybox.ethz.ch/index.php/s/OrwKOXHToZfxVyE
Yeast dataset	It is a comprehensive catalog of yeast genes [146].	A time series with the length 57 for 14 genes [32].	http://genome-www.stanford.edu/cellcycle/
THE	This is one part of The Tennessee Student/Teacher Achievement Ratio (STAR) experiment [2].	4,218 students (1,805 treated, 2,413 control), 7 covariates and binary treatment [73]	https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766

treatment, factual outcome, and counterfactual outcome are all generated by simulation, and the selection bias is created by removing treated children with non-white moms. The summarization of this year's challenge is in [42]. It reports and analyzes the performances of various methods, and concludes that methods flexibly model the response surface (i.e., the potential outcome) dominate the methods that fail to do so in terms of performance.

2017 Challenge: ACIC 2017 challenge focused on the estimation and inference for CATEs in the presence of targeted selection. Targeted selection means the likelihood that an individual receives treatment is a function of the expected response of that individual if left untreated, which leads to strong confounding [55].

2018 Challenge: The ACIC 2018 challenge has two different tasks focusing on two sub-challenges: censoring and scaling. Censoring means some of the samples may not have observed outcomes. Therefore, the dataset used by censoring challenges contains missing outcome values for some of the samples.

2019 Challenge: This challenge focuses on estimating the ATE on the quasi real-world dataset with low dimensional data and high dimensional data.[5]

IBM causal inference benchmark. This dataset is created in [143]. This dataset uses the cohort of 100K samples in Linked Births and Infant Deaths Database[6] as the fundamental set of covariates. The treatment, factual outcome, and the counterfactual outcome are generated by simulation.

BlogCatalog. This dataset is used for causal inference with networked observational data [54]. It is a social blogger network. A blogger is one observation. The bloggers are connected by some social relationships in this dataset. The features are bag-of-words representations of keywords in bloggers' descriptions. The outcomes are the opinions of readers on each blogger. A blogger belongs to the treated group (control group) if her blogs are read more on mobile devices (desktops).

Flickr. This dataset includes networked observational data in [54]. Flickr is a photo-sharing platform and social network where users upload photos for others to see. In this dataset, the users with Flickr account are observations, and the users are connected by some social relationships. The features of each user are the tags of interest. The outcomes and treatment assignment are the same as BlogCatalog.

News. The News benchmark includes 5,000 randomly sampled news articles from the NY Times corpus. It contains the data on the opinion of media consumers on news items and was originally introduced as a benchmark for counterfactual inference in the setting with two treatment options [70]. It can be extended to multiple treatments with associated dosage parameters [135].

MVICU. The **Mechanical Ventilation in the Intensive Care Unit (MVICU)** benchmark is used to estimate individual dose-response curves for a couple of treatments with an associated dosage parameter [135]. This dataset includes patients' responses to different configurations of mechanical ventilation in the intensive care unit. The data was sourced from the publicly available MIMIC III database which documents a diverse and very large population of ICU patient stays and contains comprehensive and detailed clinical data. [130].

TCGA. The Cancer Genome Atlas (TCGA) is the world's largest and richest collection of genomic data. This dataset is used to estimate individual dose-response curves for a couple of treatments with an associated dosage parameter [135]. The TCGA project collected gene expression data from various types of cancers in 9,659 individuals [168]. The treatment options are medication, chemotherapy, and surgery. The outcome is the risk of cancer recurrence after receiving the treatment.

[5]https://sites.google.com/view/acic2019datachallenge/data-challenge?authuser=0.
[6]https://www.cdc.gov/nchs/nvss/linked-birth.htm.

Table 3. Available Tool-boxes for Causal Inference

Tool-box	Supporting methods	Language	Link
Dowhy [141]	Propensity-based Stratification, PSM, IPW, Regression	Python	https://github.com/microsoft/dowhy
Causal ML [30]	Tree-based algorithms, X/T/X/R-learner	Python	https://github.com/uber/causalml
EconML [116]	Doubly Robust Learner, Orthogonal Random Forests, Meta-Learners, Deep Instrumental Variables	Python	https://github.com/microsoft/EconML#blogs-and-publications
causalToolbox	BART, Causal Forest, T/X/S-learner with BART/RF as base learner	R	https://github.com/soerenkuenzel/causalToolbox

Saccharomyces cerevisiae (yeast) cell cycle gene expression dataset. This is one time series dataset. A time series with the length T = 57 was created by combining four short time series that were measured in different microarray experiments [32].

THE. The Tennessee Student/Teacher Achievement Ratio experiment is a 4-year longitudinal class-size study funded by the Tennessee General Assembly and conducted by the State Department of Education to measure the influence of class size (small class, regular class, and regular-with-aide class) on student achievement tests and non-achievement measures [2]. Because this is one randomized controlled experiment, CATE estimates are unbiased due to unconfoundedness. Confounders are artificially introduced by selectively removing a biased subset of samples [73].

5.2 Codes/Packages

In this part, we summarize the available codes or tool-boxes for causal inference. The codes for methods that mentioned in Section 3 are provided in Tables 3 and 4, where Table 3 lists the tool-boxes with their supported methods and languages, and Table 4 lists the open-source code of one specific method.

6 APPLICATIONS

Causal inference has a variety of applications in real-world scenarios. In general, the applications of causal inference can be categorized into three directions:

(1) Decision evaluation. This is a natural application of treatment effect estimation as it is consistent with the objective of treatment effect estimation.
(2) Counterfactual estimation. Counterfactual learning greatly helps the areas related to decision making, as it can provide the potential outcomes of different decision choices (or policies).
(3) Dealing with selection bias. In many real-world applications, the records appear in the collected dataset are not representative of the whole population that is interested. Without appropriately handling the selection bias, the generalization of the trained model would be hurt.

In this section, we will discuss how causal inference benefits various real-world applications in detail.

Table 4. Available Codes of Methods in Section 3

Method	Language	Link
IPW	R	https://cran.r-project.org/web/packages/ipw/index.html
DR	R	fastDR: https://github.com/gregridgeway/fastDR DR for High dimension: https://github.com/gregridgeway/fastDR
Principal Stratification	R	https://cran.r-project.org/web/packages/sensitivityPStrat/index.html
Stratification	R	https://cran.r-project.org/web/packages/stratification/
PSM	Python	https://github.com/akelleh/causality
Perfect Match	Python	https://github.com/d909b/perfect_match
optimal matching	R	https://cran.r-project.org/web/packages/Matching/
CEM	R	https://cran.r-project.org/web/packages/cem/
TMLE	R	https://cran.r-project.org/web/packages/tmle/index.html
PSM overlap weight trapezoidal weight	Python	https://cran.r-project.org/web/packages/PSW/
Matching based Alg.: exact matching, full matching, genetic matching, nearest neighbor matching, optimal matching, subclassification	R	https://cran.r-project.org/web/packages/MatchIt/
CMGP [4]	Python	https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/ baa0aa33a6af3fe490484c9e11e3a158968ae56a/ alg/causal_multitask_gaussian_processes_ite/
BART	R Python	https://cran.r-project.org/web/packages/BayesTree/index.html https://github.com/JakeColtman/bartpy
CMGP [4]	Python	https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/ 4fb84b06c83b7ed80b681c9b7d91e66c78495378/alg/ causal_multitask_gaussian_processes_ite/
GANITE [174]	Python	https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/ baa0aa33a6af3fe490484c9e11e3a158968ae56a/alg/ganite/
BNN [70], CFR-MMD [139], CFR-WASS [139]	Python	https://github.com/clinicalml/cfrnet
CEVAE [93]	Python	https://github.com/AMLab-Amsterdam/CEVAE
SITE [171]	Python	https://github.com/Osier-Yi/SITE
grf	R	https://cran.r-project.org/web/packages/grf/index.html
R-learner	R	https://github.com/xnie/rlearner/blob/master/R/xlearner.R
Residual Balancing	R	https://github.com/swager/balanceHD
CBPS	R	https://github.com/kosukeimai/CBPS
dragonnet	Python	github.com/claudiashi57/dragonnet
Entropy Balancing	R	https://cran.r-project.org/web/packages/ebal/
DRNets [135]	Python	https://github.com/d909b/drnet

(Continued)

Table 4. Continued

Method	Language	Link
Network Deconfounder [54]	Python	https://github.com/rguo12/network-deconfounder-wsdm20
Network Embeddings [159]	Python	https://github.com/vveitch/causal-network-embeddings
RMSN [91]	Python	https://github.com/vveitch/causal-network-embeddings
TMLE [109]	R	https://github.com/joshuaschwab/ltmle
LCVA [113]	Python	https://github.com/rguo12/CIKM18-LCVA

6.1 Advertising

Properly measuring the effect of an advertising campaign can answer critical marketing questions such as whether a new advertisement increases the clicks or whether a new campaign increases sales. Since conducting randomized experiments is expensive and time-consuming, estimating the advertisement effect from the observational data is attracting increasing attention in both industry and research communities [152, 163]. In [90], the randomized NNM method is proposed to estimate the treatment effect of digital marketing campaigns. In [47], the CBGPS, discussed in Section 3.1.1, is applied to analyze the efficacy of political advertisements.

However, in the online advertisement area, it is often required to deal with complex advertisement treatments, which could be a discrete or continuous, uni- or multi-dimensional treatment [152]. One advertisement is set as the baseline treatment, and the treatment effect is obtained by comparing the potential outcome of the treatment with different values and the baseline treatment. To estimate the potential outcome of treatment with multi-dimensional values, tree-based method [163] and sparse additive model based method [152] are proposed to enable the comparison between potential treatments and the baseline treatment.

In addition to purely observational data, in the real-world scenarios, it is often the case that the dataset is comprised of large samples from control condition(i.e., the old treatment) and small samples (possibility unrepresentative) from a randomized trial which contains both the control condition and the new treatment. In [125], the small randomized trial dataset is connected with the large control dataset using the minimal set of modeling assumptions, which implies the models to predict the control and treated outcome to be similar. Under this assumption, the proposed method jointly learns the control and treated outcome predictor and regularizes the difference between the parameters of two predictors.

The above discussions show the potential applications of the treatment effect estimation in decision evaluation: measuring the effect of the advertisement campaign. Another important application is to handle the selection bias. Due to the existing selection mechanism in the advertising systems, there is a distribution discrepancy between the displayed and non-displayed events [175]. Ignoring such bias would make the advertisement click prediction inaccurate, which would cause a loss of revenue. To handle the selection bias, similar to the DR estimation mentioned in Section 3.1.1, DR policy learning is proposed in [43]. It contains two sub-estimators: direct method estimator obtained from the observed samples, and IPS estimator with the propensity score as the sample weight.

Furthermore, some works notice the difficulty of propensity score estimation due to the deterministic advertisement display policy in commercial advertisement systems. If the display policy is stochastic, the advertisements with low propensity scores still have a chance to appear in the observational dataset so that IPS can correct the selection bias. However, when the display policy is deterministic, the advertisements with low propensity scores are always absent in the observation, which makes propensity score estimation failed. This challenges motives the work of propensity-free DR method proposed in [175] which improves the original DR method in two folds: (1) train

the direct method on a small but unbiased data obtained under the uniform policy, which, to a certain degree, prevents the selection bias from propagating to the non-displayed advertisements. (2) Avoid the propensity score estimation by setting the propensity score of the observed items as 1 and combines IPS with the direct method. In a nutshell, this propensity score free method relies on the direct method trained on a small unbiased dataset to give an unbiased prediction of the advertisement click.

Apart from the applications discussed above, another important application is the advertisement recommendation, which is merged into the next subsection.

6.2 Recommendation

The recommendation is highly correlated with the treatment effect estimation, as exposing the user to an item in the recommendation system can be viewed as applying one specific treatment to a unit [83, 134]. Similar to the dataset used in the treatment effect estimation, the dataset used in the recommendation are usually biased due to the self-selection of the users. For example, in the movie rate dataset, users tend to rate the movies that they like: the horrible movie ratings are mostly made by horror movie fans and less by romantics movie fans. Another example is the advertisement recommendation datasets. The recommendation system would only recommend the advertisements to the users whom the system believes are interested in those advertisements. In the above examples, the records in the datasets are not representative of the whole population, which is the selection bias. The selection bias brings challenges to both recommendation model training and evaluation. Re-weighting samples based on the propensity score is a powerful method to solve the problems that stem from selection bias. The improved performance estimation after propensity score weighting can be calculated as follows:

$$\hat{R}_{\text{IPS}}(\hat{Y}|P) = \frac{1}{U \cdot I} \sum_{(u,i):O_{u,i}=1} \frac{\delta_{u,i}(Y, \hat{Y})}{P_{u,i}}, \tag{33}$$

where $\hat{Y}$ is the value upon which to measure the quality of a recommendation system, U is the number of users, and I is the number of items. $O_{u,i}$ is a binary variable to indicates the interaction of the u-th user with the i-th item in the observational data. $\delta_{u,i}(\cdot,\cdot)$ can be any classical quality measure of a recommendation, such as cumulative gain, discounted cumulative gain, and precision at k. P is the marginal probability matrix, whose entry is defined as $P_{u,i} = P(O_{u,i} = 1)$. The improved quality measure is an unbiased estimation to the real measurement $R(\hat{Y})$ over the whole population, which is defined as $R(\hat{Y}) = \frac{1}{U \cdot I} \sum_{u=1}^{U} \sum_{i=1}^{I} \delta_{u,i}(Y, \hat{Y})$. Based on the unbiased quality measurement, in [134], the propensity-score empirical risk minimization for recommendation is proposed: $\hat{Y} \in \mathcal{H}$ is selected to optimize the following problem: $\hat{Y}^{ERM} = \text{argmin}_{\hat{Y} \in \mathcal{H}} \{\hat{R}_{\text{IPS}}(\hat{Y}|P)\}$, where $\hat{R}_{\text{IPS}}(\hat{Y}|P)$ is defined in Equation (33). Later, various works are developed to drawbacks of propensity score weighting, including estimation variance reduction [131, 155], handle data sparsity [131, 155], DR estimation [165, 177].

In addition to using IPS or DR estimation based method to overcome the selection bias, similar to the advertisement domain, some works also adopt a small unbiased dataset to correct selection bias. In this case, the dataset contains a large set of logged feedback records under control policy and a small set of records under the randomized recommendation. CausalEmbed [22] is a representative method in this direction, which proposed a new matrix factorization algorithm. In detail, CausalEmbed jointly factorizes the matrix of those two datasets and links these two models by regularizing the difference between treated and control representation.

6.3 Medicine

Learning the optimal per-patient treatment rules is one of the promising goals of applying treatment effect estimation methods in the medical domain. When the effect of different available medicines can be estimated, doctors can give a better prescription accordingly. In [138], two challenges are mentioned to fulfill such goal: the existence of confounders and the existence of unobserved confounders. Although analyzing from the randomized experimental dataset is the golden solution, it has the following limitations: (1) the goal of randomized experimental data is to analyze the ATE instead of ITE, so the data size is often small, which limits the capacity to derive personalized treatment rules. (2) As mentioned in Section 2, conducting randomized trials is often expensive, time-consuming, and sometimes maybe unethical. Therefore, deriving personalized treatment rules from the observational dataset or the combination of experimental and observational data are two fruitful directions [138].

For the direction of utilizing an observational dataset, various methods derive the personalized treatment rules guided by the estimated ITE under the unconfoundedness assumption, such as deep-treat [9], tiered case-cohort design based method [77]. However, in this area, there are limited works to handle the unobserved confounders, and methods discussed in Section 4.2 have great potentials to explore.

6.4 Reinforcement Learning

From the perspective of reinforcement learning, ITE estimation can be viewed as a contextual multi-armed bandit problem with the treatment as the action, the outcome as the reward, and the background variables as the contextual information. Arm exploration and exploitation is similar to randomized trials and observational data. Therefore, these two areas share some similar critical challenges: (1) How to get an unbiased outcome/reward estimation? (2) How to handle either the observed or unobserved confounders that affect both the treatment assignment/action choice and the outcome/reward?

To obtain an unbiased reward estimation, importance sampling weighting [111] is the common method adopted in the offline policy evaluation. The weight is set as the probability between the target policy and the logged (observed) policy, which is analogous to IPW mentioned in Section 3.1.1. However, importance sampling proposed in [111] suffers from high variance and highly relies on the assigned weights. To improve this, similar to the DR method in ATE estimation, a DR policy evaluation is proposed in [43]. Later, various methods [10, 19, 74, 88, 154, 155, 158, 179] are proposed to improve those two methods with different settings.

As mentioned above, the second challenge is how to deal with confounders. When all the confounders are observed, we can directly optimize the unbiased reward function mentioned in the previous paragraph. However, when there exist unobserved confounders, it can lead to policies that introduce harm rather than benefit, as is generally the case with observational data [75]. The confounding-robust policy learning framework is proposed in [75], optimizing the policy over an uncertain set for propensity weights so that the unobserved confounders can be controlled.

6.5 Natural Language Processing

Recently, researchers have paid increasing attention to enhance natural language processing tasks with causal inference. De-confounded lexical feature learning [112] aims at learning the lexical features that are predictive to a set of target variables yet uncorrelated to a set of confounding variables. It has shown the importance for the linguistic analysis and the interpretation of the downstream tasks. Moreover, a series of methods have been proposed [44, 156, 169, 173] to handle the case when the treatment, confounder, or outcome contains textual data. A recent survey [76]

reviews methods and applications with textual data as the confounder. Specifically, the survey points out that textual data can serve as the confounder in two ways: one is that the text is viewed as the surrogate for potential confounders; the other way is that the language of the text itself, such as the linguistic content, could be a confounder. This survey also discusses several open problems, including mitigating the approximating error, how to design meaningful text representation, and so on.

6.6 Computer Vision

In computer vision, some emerging tasks that lie in the interaction between vision and language, such as **visual question answering (VQA)** and image captioning, have become a fundamental block for many frontier interactive AI systems [100]. Most of the existing works rely on the correlation instead of the causal relevant evidence, and the spurious correlation in the data lead to the model notoriously brittle to linguistic variations in the questions [115, 137]. To this end, a data augmentation strategy is proposed in [3], which makes the model more robust against the spurious correlation. In detail, when removing the objects irrelevant to answering the question, the model's prediction is expected to be unchanged, which prevents the model to rely on spurious correlation. When editing the objects related to the question, such as change the number of the counted objects in counting problem pairs, the model is expected to change the answer accordingly. Editing the relevant objects encourages the model to make a prediction based on the causal relevant objects. The generated complementary dataset, which includes the edited image & answering pair and the original dataset, makes the model sticking to the causally related evidence. Another line to handle the spurious correlation is to adopt the causal intervention into the model design. In [164], the authors argue that the commonsense knowledge should be included in the visual feature, but the commonsense might be confounded by the spurious observation bias. For example, the words `mouse` and `keyboard` have high co-occurrence frequency due to the underlying commonsense that they are both part of `computer`, but commonsense usually wrongly attributed to `table` because of the observation bias. To reduce such observation bias, the Visual Commonsense R-CNN is proposed which considers the object category as the confounder and directly maximizes the likelihood after the intervention to learn the visual feature representation. By removing the observation bias, the learned visual feature benefits the downstream tasks such as image captioning and VQA.

6.7 Other Applications

The applications of causal inference are not limited to the areas mentioned above, and areas related to effectiveness measurement, decision making, or handling selection bias or confounders, are all potential applications.

Education. In the education area, by comparing the outcome of different teaching methods on the student population, a better teaching method can be decided. Moreover, ITE estimation can enhance personalized learning by estimating the outcome of each student on different teaching methods. For example, ITE estimation is developed to answer the questions "Would this particular student benefit more from the video hint or the text hint when this student cannot solve a problem?", so that an Intelligent Tutor System can decide which hint is more suitable for a specific student [178].

Political decision. In the politics area, causal inference can provide decision support. For example, various methods [70, 136, 139, 171, 174] have been developed on the jobs dataset aiming to answer the question "who would benefit most from subsidized job training?". Causal inference can also help with political decisions, such as whether a policy should be promoted to a large population size.

Improving Machine learning methods. In addition to the decision support, various balancing methods that can handle the selection bias (mentioned in Section 3), can also be extended to improve the stability of machine learning methods. In [78], the reweighting method is adopted to improve the generalization ability of learned models for unknown environments (i.e., unknown test data). To be specific, the weight of each sample is added to the prediction loss function as a regularization, which is formulated as: $\sum_{j=1}^{p} || \frac{\phi(X_{.,-j}^T)(R \odot X_{.,j})}{R^T X_{.,j}} - \frac{\phi(X_{.,-j}^T)(R \odot (1-X_{.,j}))}{R^T (1-X_{.,j})} ||_2^2$, where p is the number of total features, $\phi(\cdot)$ is the feature transformation function such as neural network, $X_{.,j}$ is the j-th feature in X, $X_{.,-j}$ is the features in X except the j-th feature, $R \in \mathcal{R}^N$ is the global sample weights with N as the number of total samples. This balancing regularizer extends the CBPS method discussed in 3.1.1, by taking the j-th feature as the treatment and the remaining features as the background variables, and then combining all the features to obtain the global balancing weight.

7 FUTURE DIRECTIONS

As discussed in previous sections, existing work has made great contributions to the development of causal inference. However, there still remain a lot of open problems regarding *causal modeling and theoretical study* and *applications and evaluations*. In this section, we discuss future research directions as well as potential applications.

For causal modeling and theoretical study, we introduce several open problems as follows.

- *Adding or relaxing the assumptions in the causal model.* For instance, most of the existing approaches consider the binary treatment and the high-dimensional treatment, while the more practical settings with multiple treatments at various levels are often ignored. High-dimensional treatment is commonly observed in real life. Studying the causal interaction is a trending topic of high-dimensional treatment, which aims to identify the combinations of treatments that induce large additional effects beyond the sum of effects separately attributable to each treatment [45].
- *Developing formal connections between different causal models.* Although existing frameworks are logically the same, they have their own advantages. Building connections between different causal models benefits causal modeling from observational data. For instance, the relevance between the potential outcome framework and graphical causal models has been discussed in [67].
- *"Machine learning for causal inference" and "Causal inference for machine learning".* Machine learning and causal inference can enhance each other. Machine learning brings powerful algorithms for causal effect estimation, which is the focus of this survey. How causal inference can help improve the machine learning algorithm design, such as robustness, generalization, and knowledge transfer, is still an open problem.
- *Equip machine learning with causal reasoning capabilities.* Most of the machine learning algorithms model the correlation between variables but have very limited causal reasoning capabilities. Developing causality-aware machine learning models will help reveal the underlying mechanisms in complex observational data, and therefore assist the causality-aware predictive analysis and decision making.
- *Causal inference in dynamic environment.* Existing work mainly focuses on static observational data. In practice, data are often continuously collected from a dynamic environment. Novel causal inference approaches are required to model the dynamic observational data, leading to lifelong causal inference.

— *Causality-assisted trustworthy learning*, such as explainability, reliability, and fairness. In the model explanation domain, the causal inference has a great potential to explore the effect of the attributes to the model predicted labels. Moreover, in the fairness area, counterfactual fairness [82] is a trending topic which targets a unit's outcome in the real world and the counterfactual world where he/she has different sensitive attribute values.

Along with the rapid development of causal modeling, it is equally important to explore novel applications and build benchmarks for evaluations.

— *Generalized interpretation of "treatment" and "potential outcome" in more domain applications.* A successful example mentioned in the previous section is the recommendation system, where exposing the user to one item is analogous to applying the treatment on the unit. To expand the scope of causal inference applications, generalizing the interpretation of "treatment" and "potential outcome" in more domains is a must.
— *Integration of (partial) experimental study and observational study.* In real-world applications, sometimes, the experimental data are available, such as the A/B testing data in the web development area. Integrating the experimental data, even small sample-sized experiment data, is of great help for observational studies to overcome the unobserved confounders, and to correct the biased causal effect estimation model.
— *Extensible causal models for multi-modal data.* Multi-modal data is common in real-world applications. For instance, in the health-care domain, the doctors' records are text data and the fMRI data are images. Most of the existing treatment effect estimation models focus on one type of data, which could not handle multi-modal data. Estimating treatment effects based on multi-modal data is still an open problem.

8 CONCLUSION

The causal inference has been an attractive research topic for a long time as it provides an effective way to uncover causal relationships in real-world problems. Nowadays, the flourishing of machine learning brings new vitality into this area, and meanwhile, the incisive ideas in the causal inference area promote the development of machine learning. In this survey, we provide a comprehensive review of the methods under the well-known potential outcome framework. As the potential outcome framework relies on the three assumptions, the methods are separated into two categories. One category relies on those assumptions, while the other one relaxes some of the assumptions. For each category, we provide thorough discussions, comparisons, and summarization of the reviewed methods. The available benchmark datasets and open-source codes of those methods are also listed. Finally, some representative real-world applications of causal inference are introduced, such as advertising, recommendation, medicine, and reinforcement learning.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their valuable comments and helpful suggestions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Alberto Abadie, David Drukker, Jane Leber Herr, and Guido W. Imbens. 2004. Implementing matching estimators for average treatment effects in stata. *The Stata Journal* 4, 3 (2004), 290–311.
[2] C. M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. Tennessee's student teacher achievement ratio (STAR) project. Retrieved from http://hdl.handle.net/1902.1/10766 (2008).

[3] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 9687–9695.

[4] Ahmed M. Alaa and Mihaela van der Schaar. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 3424–3432.

[5] Ahmed M. Alaa, Michael Weisz, and Mihaela van der Schaar. 2017. Deep counterfactual networks with propensity-dropout. *CoRR* abs/1706.05966 (2017). arxiv:1706.05966 http://arxiv.org/abs/1706.05966

[6] Douglas Almond, Kenneth Y. Chay, and David S. Lee. 2005. The costs of low birth weight. *The Quarterly Journal of Economics* 120, 3 (2005), 1031–1083.

[7] Naomi Altman and Martin Krzywinski. 2015. Points of significance: Association, correlation and causation. *Nature Methods* 12, 10 (2015), 899–900.

[8] Anonymous. 2021. Continual lifelong causal effect inference with real world evidence. In *Submitted to International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=IOqr2ZyXHz1

[9] Onur Atan, James Jordon, and Mihaela van der Schaar. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *32nd AAAI Conference on Artificial Intelligence*.

[10] Onur Atan, William R. Zame, and Mihaela van der Schaar. 2018. Learning optimal policies from observational data. *arXiv preprint arXiv:1802.08679* (2018).

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

[12] Susan Athey and Guido W. Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *Statistics* 1050, 5 (2015), 1–26.

[13] Peter C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46, 3 (2011), 399–424.

[14] Heejung Bang and James M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[15] John Bather. 2000. *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*. John Wiley & Sons, Inc.

[16] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*. 137–144.

[17] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[18] Ioana Bica, Ahmed M. Alaa, and Mihaela van der Schaar. 2019. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, 884–895.

[19] Alberto Bietti, Alekh Agarwal, and John Langford. 2018. Practical evaluation and optimization of contextual bandit algorithms. *Statistics* 1050 (2018), 12.

[20] Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. 2016. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7383–7390.

[21] Colin R. Blyth. 1972. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 67, 338 (1972), 364–366.

[22] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *12th ACM Conference on Recommender Systems*. 104–112.

[23] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[24] Leo Breiman. 2017. *Classification and Regression Trees*. Routledge.

[25] Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. 1992. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics* 120, 3 (1992), 350–359.

[26] Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22, 1 (2008), 31–72.

[27] Bibhas Chakraborty. 2013. *Statistical Methods for Dynamic Treatment Regimes*. Springer.

[28] Bibhas Chakraborty and Susan A Murphy. 2014. Dynamic treatment regimes. *Annual Review of Statistics and Its Application* 1 (2014), 447–464.

[29] Yale Chang and Jennifer G. Dy. 2017. Informative subspace learning for counterfactual inference. In *31st AAAI Conference on Artificial Intelligence*.

[30] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. CausalML: Python Package for Causal Machine Learning. arxiv:cs.CY/2002.11631
[31] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (2002), 507–554.
[32] Yoichi Chikahara and Akinori Fujino. 2018. Causal inference in time series via supervised learning. In *27th International Joint Conference on Artificial Intelligence*. 2042–2048.
[33] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2007. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems*. 265–272.
[34] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.
[35] Zhixuan Chu, Stephen L. Rathbun, and Sheng Li. 2020. Matching in selective and balanced representation space for treatment effects estimation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 205–214.
[36] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96, 1 (2009), 187–199.
[37] Ralph B. D'Agostino Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17, 19 (1998), 2265–2281.
[38] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2017. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics* 221, 2 (2021), 644–654.
[39] Rajeev H. Dehejia and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94, 448 (1999), 1053–1062.
[40] Rajeev H. Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84, 1 (2002), 151–161.
[41] Peng Ding, T. J. VanderWeele, and J. M. Robins. 2017. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* 104, 2 (2017), 291–302.
[42] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science* 34, 1 (2019), 43–68.
[43] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*. 1097–1104.
[44] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163* (2018).
[45] Naoki Egami and Kosuke Imai. 2019. Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association* 114, 526 (2019), 529–540.
[46] Jianqing Fan, Kosuke Imai, Han Liu, Yang Ning, and Xiaolin Yang. 2016. *Improving Covariate Balancing Propensity Score: A Doubly Robust and Efficient Approach*. Technical Report. Technical report, Princeton Univ.
[47] Christian Fong, Chad Hazlett, and Kosuke Imai. 2018. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* 12, 1 (2018), 156–177.
[48] Constantine E. Frangakis and Donald B. Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.
[49] Steven Glazerman, Dan M Levy, and David Myers. 2003. Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science* 589, 1 (2003), 63–93.
[50] I.J. Good and Y. Mittal. 1987. The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics* 15, 2 (1987), 694–711.
[51] Xing Sam Gu and Paul R. Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 2, 4 (1993), 405–420.
[52] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2018. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 75:1–75:37.
[53] Ruocheng Guo, Jundong Li, and Huan Liu. 2019. Counterfactual evaluation of treatment assignment functions with networked observational data. In *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM*. SIAM, 271–279.
[54] Ruocheng Guo, Jundong Li, and Huan Liu. 2019. Learning individual treatment effects from networked observational data. *arXiv preprint arXiv:1906.03485* (2019).
[55] P. Richard Hahn, Vincent Dorie, and Jared S. Murray. 2019. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515* (2019).
[56] P. Richard Hahn, Jared S. Murray, and Carlos Carvalho. 2017. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis* 15, 3 (2020), 965–1056.
[57] Ben B. Hansen. 2008. The prognostic analogue of the propensity score. *Biometrika* 95, 2 (2008), 481–488.

[58] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *34th International Conference on Machine Learning-Volume 70*. 1414–1423.
[59] Negar Hassanpour and Russell Greiner. 2019. Counterfactual regression with importance sampling weights. In *28th International Joint Conference on Artificial Intelligence*. 5880–5887.
[60] James J. Heckman, Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65, 2 (1998), 261–294.
[61] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
[62] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*. 689–696.
[63] Michael G. Hudgens and M. Elizabeth Halloran. 2008. Toward causal inference with interference. *Journal of the American Statistical Association* 103, 482 (2008), 832–842.
[64] Katherine Huppler Hullsiek and Thomas A. Louis. 2002. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* 3, 2 (2002), 179–193.
[65] Stefano M. Iacus, Gary King, and Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis* 20, 1 (2012), 1–24.
[66] Kosuke Imai and Marc Ratkovic. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 1 (2014), 243–263.
[67] Guido Imbens. 2019. *Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics*. Technical Report. National Bureau of Economic Research.
[68] Guido W. Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86, 1 (2004), 4–29.
[69] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
[70] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*. 3020–3029.
[71] Fredrik D. Johansson, Nathan Kallus, Uri Shalit, and David Sontag. 2018. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598* (2018).
[72] Judea Pearl. 2012. Judea Pearl on Potential Outcomes. Retrieved from http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/.
[73] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. 2018. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*. 10888–10897.
[74] Nathan Kallus and Masatoshi Uehara. 2019. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems*. 3320–3329.
[75] Nathan Kallus and Angela Zhou. 2018. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*. 9269–9279.
[76] Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5332–5344.
[77] Ronald C. Kessler, Robert M. Bossarte, Alex Luedtke, Alan M. Zaslavsky, and Jose R. Zubizarreta. 2019. Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy* 120 (2019), 103412.
[78] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable prediction across unknown environments. In *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1617–1626.
[79] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating treatment effect in the wild via differentiated confounder balancing. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 265–274.
[80] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. 2017. Treatment effect estimation with data-driven variable decomposition. In *31st AAAI Conference on Artificial Intelligence*.
[81] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.
[82] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[83] Akos Lada, Alexander Peysakhovich, Diego Aparicio, and Michael Bailey. 2019. Observational data for heterogeneous treatment effects with application to recommender systems. In *2019 ACM Conference on Economics and Computation*. 199–213.

[84] Philip W. Lavori and Ree Dawson. 2000. A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163, 1 (2000), 29–38.
[85] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. 2011. Weight trimming and propensity score weighting. *PloS One* 6, 3 (2011), e18174.
[86] Changhee Lee, Nicholas Mastronarde, and Mihaela van der Schaar. 2018. Estimation of individual treatment effect in latent confounder models via adversarial learning. *arXiv preprint arXiv:1811.08943* (2018).
[87] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 113, 521 (2018), 390–400.
[88] Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. 2012. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Workshop on On-line Trading of Exploration and Exploitation 2*. 19–36.
[89] Sheng Li and Yun Fu. 2017. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*. 929–939.
[90] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *25th International Joint Conference on Artificial Intelligence*. 3768–3774.
[91] Bryan Lim. 2018. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*. 7483–7493.
[92] Wei-Yin Loh. 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, 1 (2011), 14–23.
[93] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*. 6446–6456.
[94] Xinwei Ma and Jingshen Wang. 2010. Robust inference using inverse probability weighting. *J. Amer. Statist. Assoc.* 115, 532 (2020), 1851–1860.
[95] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation: Learning bounds and algorithms. In *The 22nd Conference on Learning Theory*.
[96] Susan A. Murphy. 2003. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65, 2 (2003), 331–355.
[97] Susan A. Murphy. 2005. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24, 10 (2005), 1455–1481.
[98] Jessica A. Myers, Jeremy A. Rassen, Joshua J. Gagne, Krista F. Huybrechts, Sebastian Schneeweiss, Kenneth J. Rothman, Marshall M. Joffe, and Robert J. Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174, 11 (2011), 1213–1222.
[99] Xinkun Nie and Stefan Wager. 2017. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912* (2017).
[100] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2020. Counterfactual VQA: A cause-effect look at language bias. *CoRR* abs/2006.04315 (2020). arxiv:2006.04315 https://arxiv.org/abs/2006.04315
[101] Elizabeth L. Ogburn and Tyler J. VanderWeele. 2014. Causal diagrams for interference. *Statistical Science* 29, 4 (2014), 559–578.
[102] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[103] Judea Pearl. 2000. *Causality: Models, Reasoning and Inference*. Vol. 29. Springer.
[104] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3 (2009), 96–146.
[105] Judea Pearl. 2009. *Causality*. Cambridge University Press.
[106] Judea Pearl. 2012. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. 417–424.
[107] Judea Pearl. 2014. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.
[108] Jose M. Peña. 2018. Reasoning with alternative acyclic directed mixed graphs. *Behaviormetrika* 45, 2 (2018), 389–422.
[109] Maya Petersen, Joshua Schwab, Susan Gruber, Nello Blaser, Michael Schomaker, and Mark van der Laan. 2014. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference* 2, 2 (2014), 147–185.
[110] Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. 2012. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21, 1 (2012), 31–54.
[111] Doina Precup. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* (2000), 80.
[112] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded lexicon induction for interpretable social science. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1615–1625.

[113] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *27th ACM International Conference on Information and Knowledge Management*. 1679–1682.

[114] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics* 3, 2 (2017), 121–129.

[115] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! Improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5863–5868.

[116] Microsoft Research. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. Retrieved from https://github.com/microsoft/EconML. Version 0.x.

[117] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* 22, 4 (2007), 544–559.

[118] James M. Robins. 2004. Optimal structural nested models for optimal sequential decisions. In *2nd Seattle Symposium in Biostatistics*. Springer, 189–326.

[119] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 427 (1994), 846–866.

[120] Peter M. Robinson. 1988. Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society* 56, 4 (1988), 931–954.

[121] Paul R. Rosenbaum. 1987. Model-based direct adjustment. *Journal of the American Statistical Association* 82, 398 (1987), 387–394.

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

[123] Paul R. Rosenbaum and Donald B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79, 387 (1984), 516–524.

[124] Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1 (1985), 33–38.

[125] Nir Rosenfeld, Yishay Mansour, and Elad Yom-Tov. 2017. Predicting counterfactuals from large historical data and small randomized trials. In *26th International Conference on World Wide Web Companion*. 602–609.

[126] Donald B. Rubin. 1973. Matching to remove bias in observational studies. *Biometrics* 29, 1 (1973), 159–183.

[127] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies.*Journal of Educational Psychology* 66, 5 (1974), 688.

[128] Donald B. Rubin and Neal Thomas. 1996. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52, 1 (1996), 249–264.

[129] Donald B. Rubin and Neal Thomas. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95, 450 (2000), 573–585.

[130] Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine* 39, 5 (2011), 952.

[131] Yuta Saito. 2019. Eliminating bias in recommender systems via pseudo-labeling. *arXiv preprint arXiv:1910.01444* (2019).

[132] Brian C. Sauer, M. Alan Brookhart, Jason Roy, and Tyler VanderWeele. 2013. A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and Drug Safety* 22, 11 (2013), 1139–1145.

[133] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. 1999. Comments and rejoinder. *Journal of the American Statistical Association* 94, 448 (1999), 1121–1146.

[134] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*. 1670–1679.

[135] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. 2020. Learning counterfactual representations for estimating individual dose-response curves. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 5612–5619.

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

[137] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6649–6658.
[138] Uri Shalit. 2019. Can we learn individual-level treatment policies from clinical data? *Biostatistics* 21, 2 (2019), 359–362. DOI: https://doi.org/10.1093/biostatistics/kxz043
[139] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *34th International Conference on Machine Learning-Volume 70*. 3076–3085.
[140] Cosma Rohilla Shalizi and Andrew C. Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40, 2 (2011), 211–239.
[141] Amit Sharma and Emre Kiciman. 2019. DoWhy: A Python package for causal inference. Retrieved from https://github.com/microsoft/dowhy.
[142] Eli Sherman and Ilya Shpitser. 2018. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems*. 9424–9435.
[143] Y. Shimoni, C. Yanover, E. Karavani, and Y. Goldschmnidt. 2018. Benchmarking framework for performance-evaluation of causal inference analysis. *ArXiv preprint arXiv:1802.05046* (2018).
[144] Ilya Shpitser. 2015. Segregated graphs and marginals of chain graph models. In *Advances in Neural Information Processing Systems*. 1720–1728.
[145] Jeffrey Smith. 2000. *A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies*. Technical Report.
[146] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. 1998. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* 9, 12 (1998), 3273–3297.
[147] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*. MIT Press.
[148] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. In *Applied Informatics*, Vol. 3. Springer, 3.
[149] Jerzy Splawa-Neyman, Dorota M. Dabrowska, and T. P. Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9.*Statistical Science* 5, 4 (1990), 465–472.
[150] Morgan Stephen and Winship Christopher. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge, UK.
[151] Elizabeth A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.
[152] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. 2015. Causal inference via sparse additive models with application to online advertising. In *29th AAAI Conference on Artificial Intelligence*. 297–303.
[153] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 3104–3112.
[154] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*. 814–823.
[155] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. 2017. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*. 3632–3642.
[156] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. The Association for Computer Linguistics, 175–185.
[157] Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. 2012. On causal inference in the presence of interference. *Statistical Methods in Medical Research* 21, 1 (2012), 55–75.
[158] Guy Tennenholtz, Shie Mannor, and Uri Shalit. 2020. Off-Policy Evaluation in Partially Observable Environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10276–10283.
[159] Victor Veitch, Yixin Wang, and David Blei. 2019. Using embeddings to correct for unobserved confounding in networks. In *Advances in Neural Information Processing Systems*. 13769–13779.
[160] Thomas Verma and Judea Pearl. 1991. *Equivalence and Synthesis of Causal Models*. UCLA, Computer Science Department.
[161] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, A Rusu Andrei, and Veness Joel. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.
[162] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113, 523 (2018), 1228–1242. DOI: https://doi.org/10.1080/01621459.2017.1319839

[163] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. 2015. Robust tree-based causal inference for complex ad effectiveness analysis. In *8th ACM International Conference on Web Search and Data Mining*. 67–76.
[164] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense representation learning via causal inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 378–379.
[165] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*. 6638–6647.
[166] Christopher Watkins. 1989. *Learning From Delayed Rewards*. Ph.D. Dissertation. King's College, Cambridge.
[167] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8, 3–4 (1992), 279–292.
[168] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, Cancer Genome Atlas Research Network, et al. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics* 45, 10 (2013), 1113.
[169] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Conference on Empirical Methods in Natural Language Processing*. NIH Public Access, 4586.
[170] Jeffrey M. Wooldridge. 2016. Should instrumental variables be used as matching variables? *Research in Economics* 70, 2 (2016), 232–237.
[171] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*. 2633–2643.
[172] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2019. ACE: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining*. 1432–1437.
[173] Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. 2019. On the estimation of treatment effect with text covariates. In *28th International Joint Conference on Artificial Intelligence*. 4106–4113.
[174] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *6th International Conference on Learning Representations*.
[175] Bo-Wen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. 2019. Improving ad click prediction by considering non-displayed events. In *28th ACM International Conference on Information and Knowledge Management*. 329–338.
[176] Kun Zhang and Aapo Hyvarinen. 2009. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 647–655.
[177] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2019. A causal perspective to unbiased conversion rate estimation on data missing not at random. *arXiv preprint arXiv:1910.09337* (2019).
[178] Siyuan Zhao and Neil Heffernan. 2017. Estimating individual treatment effects from educational studies with residual counterfactual networks. In *10th International Conference on Educational Data Mining*.
[179] Hao Zou, Kun Kuang, Boqi Chen, Peixuan Chen, and Peng Cui. 2019. Focused context balancing for robust offline policy evaluation. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 696–704.

Received April 2020; revised November 2020; accepted December 2020