Causal Inference with Complex Treatments: A Survey

YINGRONG WANG, Zhejiang University, Hangzhou, China
HAOXUAN LI, Peking University, Beijing, China
MINQIN ZHU, Zhejiang University, Hangzhou, China
ANPENG WU, Zhejiang University, Hangzhou, China
BAOHONG LI, Zhejiang University, Hangzhou, China
KETING YIN, Zhejiang University, Hangzhou, China
RUOXUAN XIONG, Emory University, Atlanta, United States
FEI WU, Zhejiang University, Hangzhou, China
KUN KUANG*, Zhejiang University, Hangzhou, China

Causal inference plays an important role in explanatory analysis and decision-making across a wide range of fields, including statistics, marketing, healthcare, and education. Its core objective is to estimate treatment effects and inform intervention policies. Most existing work focuses on the binary treatment setting, where each unit is assigned to either treatment or control. In practice, however, treatments are often more complex, encompassing multi-valued, continuous, or bundle interventions. We refer to such settings as complex treatments. In this paper, we provide a systematic and comprehensive survey of causal inference methods for complex treatments. We first revisit the problem formulation, core assumptions, and their possible variations under different settings. We sequentially review the representative methods for multi-valued, continuous, and bundle treatments. Within each setting, we organize the methods into two broad categories: those that rely on the *unconfoundedness assumption* and those that address violations of this assumption. We further discuss the intrinsic relationships among these methods and the assumption verification. Finally, we summarize available benchmark datasets and open-source codes, and outline several directions for future research.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Causal reasoning and diagnostics**.

Additional Key Words and Phrases: causal inference, multi-valued treatment, continuous treatment, bundle treatment

1 Introduction

Causal inference is revolutionizing diverse fields, from statistics to healthcare, by moving beyond mere association to pinpoint the actual impact of treatments or interventions [105]. Its extensive applications span numerous domains, including marketing [136], epidemiology [111], education [66], and even recommendation systems [141]. Unlike traditional models that rely on correlation for pattern recognition, causal methods directly tackle the complex challenge of determining causal effects–the difference in outcomes with and without a specific

*Corresponding author.

Authors' Contact Information: Yingrong Wang, Zhejiang University, Hangzhou, China; e-mail: wangyingrong@zju.edu.cn; Haoxuan Li, Peking University, Beijing, China; e-mail: hxli@stu.pku.edu.cn; Minqin Zhu, Zhejiang University, Hangzhou, China; e-mail: minqinzhu@zju.edu.cn; Anpeng Wu, Zhejiang University, Hangzhou, China; e-mail: anpwu@zju.edu.cn; Baohong Li, Zhejiang University, Hangzhou, China; e-mail: baohong.li@zju.edu.cn; Keting Yin, Zhejiang University, Hangzhou, China; e-mail: yinkt@zju.edu.cn; Ruoxuan Xiong, Emory University, Atlanta, Georgia, United States; e-mail: ruoxuan.xiong@emory.edu; Fei Wu, Zhejiang University, Hangzhou, China; e-mail: wufei@cs.zju.edu.cn; Kun Kuang, Zhejiang University, Hangzhou, China; e-mail: kunkuang@zju.edu.cn.

ACM 1557-7341/2026/1-ART
https://doi.org/10.1145/3789499

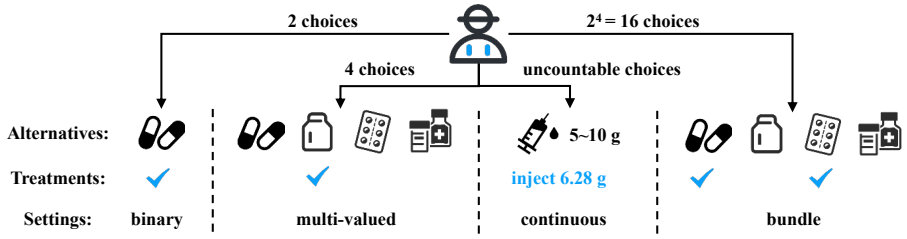


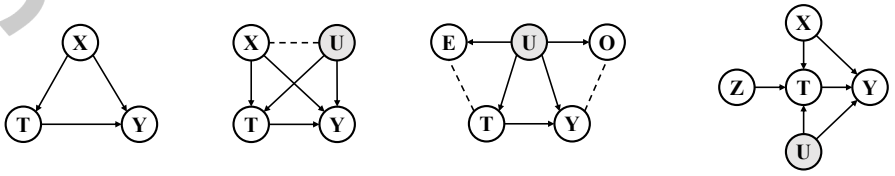
Fig. 1. An example of complex treatments.

intervention. This goes beyond simple measurement, aiding in critical downstream tasks such as prediction, decision-making, and explanatory analysis. The primary challenge is controlling confounding bias, where external factors simultaneously influence both treatment and outcome, leading to skewed estimations.

Randomized Controlled Trials (RCTs) are the gold standard for estimating treatment effects, as random assignment inherently minimizes confounding. However, RCTs are often impractical or ethically challenging, e.g., forcing patients to forgo a life-saving drug for a study. This drives the increasing focus on extracting causal insights from observational data, which is naturally collected but is inherently susceptible to confounding. Recently, a variety of methodologies have emerged to navigate the complexities of observational data, such as propensity score-based methods [113] (e.g., matching [27] and re-weighting [114]), representation learning methods [62] (e.g., CounterFactual Regression (CFR) [122] and Dragonnet [124]), and generative modeling approaches (e.g., GANITE [157] and CEVAE [83]). However, these approaches, although effective, are mainly focused on binary treatments and cannot directly address the nuances of more complex interventions.

In practical applications, causal inference with complex treatments has attracted increasing interest, where the treatment could be multi-valued, multi-dimension (bundle), continuous, or even more complex. For example, as shown in Fig. 1, in healthcare, when choosing among different alternatives for a treatment, the treatments themselves can vary significantly in their settings. In a "binary" setting, there are only 2 choices, such as taking a capsule or not. A "multi-valued" setting offers 4 choices, perhaps different types of medication (e.g., a capsule, a liquid, or a blister pack) or combinations thereof. A "continuous" setting might involve a dosage range, like injecting 5 to 10 grams of a substance, which offers uncountable choices within that range, exemplified by a precise dose of "inject 6.28 g". Finally, a "bundle" setting represents a combination of multiple treatments, leading to $2^4 = 16$ possible choices if there are four distinct treatment components that can be chosen or not. These examples highlight the need for causal inference methods that can handle such diverse and intricate treatment structures beyond simple binary interventions.

Another significant challenge in complex treatment is the presence of unmeasured confounding variables. Due to the lack of standardized data collection protocols and user privacy issues, some key variables are often unobservable in real-world datasets. We refer to them as “unobserved confounders” that are typically common causes of both treatment T and outcome Y, as illustrated in Fig. 2(a) and Fig. 2(b). When such unobserved confounders (U) are present, even after controlling for observed variables (X), systemic biases stemming from



(a) Observed (b) Unobserved (c) Proxy Variable (d) Instrumental Variable

Fig. 2. Potential outcome frameworks w/o unobserved confounders (marked in shadow).

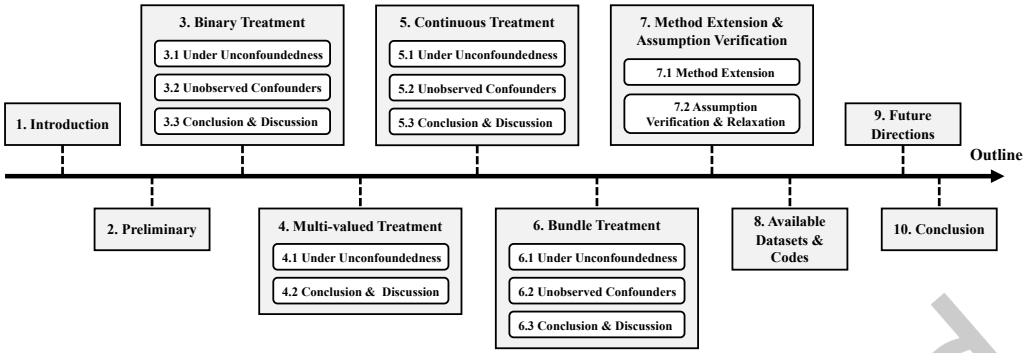


Fig. 3. Outline of the survey.

these hidden factors can substantially undermine the accuracy of treatment effect estimation. This inherent bias fundamentally limits the ability to accurately estimate causal effects under unmeasured confounding.

A promising approach to address the challenge of U is to use proxy variables, which are observable and can help capture the underlying influence of U. Proxy variables serve as indirect indicators of the effect from U, allowing us to approximate and control these latent factors in the analysis. Taking negative control as an example, as illustrated in Fig. 2(c), negative control exposures E and negative control outcomes O serve as proxy variables that can assist in addressing unobserved confounding. Negative controls are particularly useful because they offer a way to test and adjust for hidden biases when direct measurement of confounders is not possible. Relevant works include Multiple Causal Estimation via Information (MCEI) [108] for multi-valued treatment, Deep Feature Proxy Variable (DFPV) [154] and Identifiable treatment-conditional VAE (Intact-VAE) [150] for continuous treatment, and Task Embedding based Causal Effect VAE (TECE-VAE) [116] for bundle treatment. In addition, instrumental variable (IV) is also widely used, which is illustrated in Fig. 2(d). Given X, the instrumental variable Z is beneficial for the identification of $T \rightarrow Y$. DeepIV [45] and Kernel IV [127] are two examples of such instrumental variable methods.

There are several surveys in the causal inference community, such as the two focused on binary treatment [39, 156], the work that concludes instrumental variable methods [147], and the one discussing matching methods for multi-valued treatment [82]. However, the problem of estimating causal effect of complex treatments is rarely discussed, which is common and important in practical applications. In this paper, we provide a comprehensive review of methods with complex treatments under the potential outcome framework. Paper organization is illustrated in Fig. 3.

2 Preliminary

We first introduce the basic setups in the case of binary treatment. Suppose there is a random sample of n units from a population $\mathcal{P}$. Let $X_i \in \mathcal{X} \subset \mathbb{R}^d$ denote the covariate of each unit i, and $T_i \in \mathcal{T}^{bin} = \{0, 1\}$ denote the assigned treatment. When there exist m discrete treatments, we rewrite the treatment as $T_i \in \mathcal{T}^{mul} = \{0, 1, \ldots, m\}$ for the multi-valued setting, and $T_i \in \mathcal{T}^{bun} \subset \{0, 1\}^m$ for the bundle setting. As for the continuous treatment, it can be denoted as $T_i \in \mathcal{T}^{con} \subset \mathbb{R}$. The outcome of unit i receiving a specific treatment T_i is $Y_i \in \mathcal{Y} \subset \mathbb{R}$. Note that we consider the circumstance of continuous outcome in this paper. We adopt the potential outcome framework [115, 130] in causal inference. For generality, let $Y_i(t)$ and $Y_i(0)$ be the outcome of receiving treatment $T_i = t$ and no treatment $T_i = 0$. Only one of them can be observed in the dataset while the other is obtained by counterfactual prediction, which is known as the fundamental problem of causal inference [51, 95]. In Table **??**, we summarize some important notations that are commonly used.

In causal inference, the leap from observed correlations between the aforementioned variables to the causal relationships between them relies on a set of foundational assumptions. These assumptions are crucial because they define the conditions under which a statistical association can be reliably interpreted as a true causal relationship, especially when we cannot directly observe what would have happened to an individual if receiving different treatments.

(1) *Overlap* or *positivity assumption*, formally stated as $\mathbb{P}(T = t|X = x) > 0, \forall\, t, x$. It ensures that for every combination of observed X, there is a non-zero probability of receiving any given treatment ($T = t$). This requires the data to cover all treatment groups, allowing comparisons between these groups. If *overlap* is violated (e.g., a certain group *never* receives a particular treatment), then it is impossible to compare the two outcomes for causal effect.
(2) *Unconfoundedness* or *ignorability assumption*, that $Y(T = t) \perp\!\!\!\perp T \,|\, X$. It guarantees that there are no unobserved confounders U that simultaneously affect T and Y. If this assumption holds, after accounting for observed characteristics X, any remaining association can be attributed to a causal effect rather than a shared common cause.
(3) *Stable unit treatment value assumption (SUTVA)*, which contains considerations from two aspects. *No interference*: units are independent of each other and do not influence the outcomes of the others. *Treatment consistency*: there are no alternative forms of a treatment, ensuring the observed outcome truly corresponds to the assigned potential outcome.

Unconfoundedness and *overlap* are collectively referred to as *strong ignorability assumption* [56].

Based on the aforementioned counterfactual outcome and assumptions, individual treatment effect (ITE) of unit i can be measured as $ITE_i = Y_i(T_i = t) - Y_i(T_i = 0)$. As for the whole population, we use average treatment effect (ATE) to quantify the treatment effect, i.e., $ATE = \mathbb{E}[Y(T = t) - Y(T = 0)]$. If we only focus on that of the treated group (ATT), we have $ATT = \mathbb{E}[Y(T = t) - Y(T = 0)|T = t]$. To study the effect on samples with particular characteristics, we can pick out a subgroup and the effect on them is called conditional average treatment effect (CATE), i.e., $CATE = \mathbb{E}[Y(T = t) - Y(T = 0)|X = x]$. This measurement also plays an important role when treatment effect varies significantly across different subgroups, which is also known as the heterogeneous treatment effect (HTE).

The three basic assumptions and targeted estimands are generalizable across settings of binary, multi-valued and bundle treatments, because of limited treatment space (e.g., $\{2, 4, 16\}$ in Fig. 1). For bundle treatment, additional assumptions (e.g., interactions between treatments [92]) might be required since multiple treatments are simultaneously taken. When it comes to continuous treatment, an additional *smoothness assumption* is introduced, stating that the potential outcome function is assumed to be continuously differentiable with respect to the treatment variable. It ensures that there are no jump discontinuities in the dose-response function. Besides, the *unconfoundedness assumption* for continuous treatment should be rewritten as $Y(t) \perp\!\!\!\perp T \,|\, X, \forall t \in T$. The key to measuring the effects of continuous treatment is the dose-response function, and there are various measurements [31, 34]. Similarly to ITE, the formal definition of individual dose-response function (IDRF) is given as $IDRF_i(t) = Y_i(T_i = t)$. The average dose-response function (ADRF) quantifies the causal effect on the whole population by $ADRF(t) = \mathbb{E}[Y(T = t)]$. To capture the heterogeneity of continuous treatment effects, ADRF is naturally extended as the heterogeneous dose-response function (HDRF), which is formally defined as $HDRF(t, x) = \mathbb{E}[Y(T = t)|X = x]$.

3 Binary Treatment

Considering that many methods for complex treatments are developed from models of the binary setting, we first give a brief introduction of those binary treatment methodologies. This section is to serve as the support of background knowledge for the following parts.

3.1 Under Unconfoundedness

We first introduce the methods when all three basic assumptions hold. Traditional statistical methods include propensity score, outcome regression, and covariate balancing. Later, tree-based machine learning methods emerged. With the development of deep neural networks, representation learning methods and generative modeling methods have caused increasing interest among researchers.

3.1.1 Propensity Score (PS)-Based Methods. Propensity score is one of the most common methods used for binary treatment setting. Its definition can be described as the following equation, which refers to the conditional probability of a unit receiving treatment T when given covariates X:

$$e(X) = \mathbb{P}(T = 1|X). \tag{1}$$

By using methods such as matching [27], stratification [114], and re-weighting [112], we can simulate an RCT environment to make samples in the treated group and the control group similar, thus alleviating the confounding from $X \to T$. In **matching** methods, a unit i whose neighbors matched from the opposite group are denoted as $\mathcal{J}(i)$, we can estimate the counterfactual outcomes from the observed outcomes by $\hat{Y}_i(1-t) = \frac{1}{|\mathcal{J}(i)|}\sum_{j\in\mathcal{J}(i)} Y_j(t)$. Various designs on distance measurement and matching algorithm are discussed in the survey [131]. **Stratification** methods, also named *sub-classification* or *blocking*, are aimed to split the entire group into several subgroups (blocks), within each those units in the treated group and the control group are similar. The way to split all the samples is based on the propensity score as well. Suppose that there are J blocks, we can estimate $ATE_{strat} = \sum_{j=1}^{J} \frac{n_j}{n}\left[\bar{Y}_t(j) - \bar{Y}_c(j)\right]$, where n_j is the number of samples in the j-th block, and $\bar{Y}_t(j)$ and $\bar{Y}_c(j)$ are the average outcomes of the treated group and control group, respectively. **Re-weighting** methods are focused on assigning appropriate weight to each unit in order to construct a new population where distributions of the treated group and control group are similar. Taking inverse propensity weighting (IPW) [112, 113] for example, the sample weights w can be defined as $w = \frac{T}{\hat{e}(X)} + \frac{1-T}{1-\hat{e}(X)}$, where $\hat{e}$ is the estimated value of propensity score. Therefore,

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}\right]. \tag{2}$$

However, these methods are sensitive to model misspecification and extreme values of propensity scores. Taking IPW for example, a particularly small $\hat{e}$ in the denominator can result in a particularly large w, and even a slight error of $\hat{e}$ can cause a significant error of the estimated $\hat{ATE}_{IPW}$.

3.1.2 Outcome Regression Methods. Different from the PS-based methods that aim to calibrate observational data to resemble data from RCTs, outcome regression methods directly fit $\mathbb{P}(Y|X,T)$ via supervised learning and infer the counterfactual outcome by $\mathbb{E}[Y|X, 1-T]$. A basic solution is **S-learner** [71], which regards T as part of X and uses a single machine learning model $\mu(X,T)$ to directly predict $\hat{CATE} = \hat{\mu}(X,1) - \hat{\mu}(X,0)$. When T is strongly correlated with X, $\mu(\cdot)$ may struggle to capture the heterogeneity of the treatment effect. In addition, when X is of high dimension, the information from the 1-dimensional T would be neglected. Therefore, **T-learner** [71] establishes two distinct learning models $\mu_1(X) = \mathbb{E}[Y|X, T=1]$ and $\mu_0(X) = \mathbb{E}[Y|X, T=0]$, estimating $\hat{CATE} = \hat{\mu}_1(X) - \hat{\mu}_0(X)$. T-learner can perform poorly when the sample sizes of the two groups are imbalanced. To address this, **X-learner** [71] utilizes cross-group information, i.e., improving the effect estimator of treated group with information from the control group and vice versa. Specifically, following the preliminary training of $\mu_1(\cdot)$ and $\mu_0(\cdot)$, two separate CATE estimators $\{\tau_1(\cdot), \tau_0(\cdot)\}$ are learned using the derived residuals $Y_1 - \hat{\mu}_0(X)$ and $\hat{\mu}_1 - Y_0$, respectively. However, these methods do not address the problem of confounding bias. **Doubly Robust (DR)** [88], also called Augmented IPW, combines the outcome regression and propensity score as follows:

$$ATE_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}\right], \tag{3}$$

where $\hat{\mu}$ is the estimated value of an outcome regression model $\mu(\cdot)$. Compared to IPW, DR relaxes the assumption of model correctness, only requiring at least one model to be correctly specified. In this way, even if $\hat{\mu}$ or $\hat{e}$ is poorly estimated, the overall estimator is still robust. **Double Machine Learning (DML)** [21] employs two machine learning models as backbones of $e(\cdot)$ and $\mu(\cdot)$, to separately predict T and Y based on X. The residuals $\epsilon_Y = Y - \hat{\mu}(X)$ and $\epsilon_T = T - \hat{e}(X)$ are orthogonal to X. Therefore, by regressing ϵ_Y on ϵ_T, the regression coefficient serves as the estimate of ATE. **R-learner** [99] can be regarded as a form of DML, except that it jointly trains all models through $\arg\min \mathbb{E}[(Y - \hat{\mu}(X) - (T - \hat{e}(X))\hat{\tau}(X)]$, where $\hat{\tau}(X) = \frac{\epsilon_Y}{\epsilon_T}$ is the estimate of CATE.

3.1.3 Covariate Balancing Methods. Considering that the regression of PS often relies on model specification, researchers propose alternative approaches that directly adjust the covariate distribution of two groups so as to control the selection bias. It is like simulating the randomization process with observational data for the purpose of achieving $T_i \perp X_i$. The main idea is to assign weights to each sample to ensure the re-weighted groups satisfy the balance constraints, that is, aligning the first-order moment of sample covariates between the treated group and the control group. **Entropy balancing** [40] method determines the re-weighting scheme by minimizing the entropy divergence between distributions of the two groups. As for **covariate balancing propensity score (CBPS)** [53], it utilizes the balancing property of propensity score, i.e., $T_i \perp X_i | e(X_i)$, to improve its estimation. To be specific, the propensity scores are solved by $\mathbb{E}\left[\frac{T_i\tilde{X}_i}{e(X_i)} - \frac{(1-T_i)\tilde{X}_i}{1-e(X_i)}\right] = 0$, where w_i represents the weights of X_i, and $\tilde{X}_i = w_i X_i$ is the adjusted covariates after re-weighting. **Approximate residual balancing** [7] is another method that incorporates the concepts from doubly robust estimation. Specifically, it combines balancing weights learning, propensity score regression, and potential outcome estimation together. **Kernel balancing** [143] is proposed in recent years, which attains uniform approximate balance for covariate functions in a reproducing-kernel Hilbert space.

3.1.4 Tree-based Methods. The tree-based models, such as the Classification And Regression Tree (CART) [13], is also widely used to estimate heterogeneity in causal effects. We can use the tree splitting to partition the samples into several sub-groups. Afterwards, treatment effects are estimated according to other samples that fall into the same leaf as the target unit. **Bayesian Additive Regression Trees (BART)** [23, 24] is a Bayesian ensemble method modeling the mean outcome from a sum of M trees, i.e., $Y = \sum_{j=1}^{M} g(X; h_j, \theta_j) + \varepsilon$, where h_j denotes the j-th regression tree, $\theta_j = \{\mu_1, \ldots, \mu_B\}$ is a set of parameters associated with each of the B terminal nodes of h_j, mapping function $g(X; h_j, \theta_j)$ assigns the corresponding leaf value $\mu_b \in \theta_j$ to the given X, and ε is the error term following a normal distribution with a mean of zero. The estimand of interest is obtained by contrasting the imputed potential outcomes between treatment groups. The tree methods are naturally applicable to tackle the causal effect estimation of multi-valued treatment.

3.1.5 Representation-based Methods. **Balancing Neural Network (BNN)** [62] and **CounterFactual Regression (CFR)** [122] are two widely recognized methods based on invariant representation. Generally speaking, such methods include a representation network $\Phi(x)$ to learn the universal representation for samples from both groups, together with a hypothesis network $h(\Phi)$ to predict potential outcomes. Considering that the populations of different treatment groups are supposed to be similar or balanced, constraints are applied to minimize the discrepancy between their distributions. Therefore, the basic objective function of representation-based methods can be concluded as $\mathcal{L} = \mathcal{L}(h) + \mathcal{L}(\Phi) + \mathcal{R}$. The first term refers to the prediction error the hypothesis network. The second term is a quantitative measurement of the discrepancy distance between the distributions of the treated group and the control group. The last term $\mathcal{R}$ is an optional regularization term that controls model complexity. Take CFR for example, its objective function is:

$$\min_{\substack{h,\Phi \\ \|\Phi\|}} \frac{1}{n} \sum_{i=1}^{n} w_i \cdot \mathcal{L}\left(h(\Phi(X_i), T_i), Y_i\right) + \alpha \cdot \mathrm{IPM}_{\mathrm{G}}\left(\{\Phi(X_i)\}_{i:T_i=0}, \{\Phi(X_i)\}_{i:T_i=1}\right) + \lambda \cdot \mathcal{R}(h), \tag{4}$$

where $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$ and $u = \frac{1}{n}\sum_{i=1}^{n} T_i$. Loss function $\mathcal{L}$ is often modeled as mean squared error (MSE) for continuous outcomes. The distance between two distribution is measured by integral probability metric (IPM), i.e., the term denoted as IPM_G. Weights α and λ are hyper-parameters.

3.1.6 Generative Modeling Methods. **GANITE [157]** introduces the idea of Generative Adversarial Network (GAN) [36] into the causal inference community. It is composed of a counterfactual block and an ITE block, and in each block there is a separate GAN structure. In the counterfactual block, the generator is to fill up all the missing counterfactual outcomes Y_i^{CF} while the discriminator is to decide whether the potential outcome is the real data Y_i^F or the fake ones derived from the generator. In this way, a "complete" dataset can be obtained. As for the ITE block, there is a generator to estimate the outcome $\widehat{ITE}_i$ given the covariate X_i, and a discriminator aimed at distinguishing whether its input is ITE_i form the dataset after imputation or $\widehat{ITE}_i$ from the generator. Ultimately, the two generators can give accurate estimations of Y_i^{CF} and $\widehat{ITE}_i$.

3.2 With Unobserved Confounders

Proxy variable and instrumental variable are powerful tools when there exist unobserved confounders. Although these methods are originally proposed to tackle binary treatment, they can be extended to address the problem of multi-valued and continuous treatment. Therefore, we provide a brief overview of these concepts in this section, and discuss some concrete methods in Section 5.2.

3.2.1 Proxy Variable. It has been studied for a long time as bias analysis [9, 37]. The main idea of proxy is briefly described in Section 1 and Fig. 2(c). We tentatively divide the proxy variable methods into two categories, including negative controls and generative modeling methods.

Negative Controls. Many works [69, 91, 125] have introduced the concept called negative control variables. They can be understood as variables collected under a controlled experiment with negative expectations, i.e., the expected effect is not anticipated ($T \not\to Y$). Negative controls can be divided into negative control outcome (NCO) and negative control exposure (NCE). As shown in Fig. 2(c), NCO (denoted as O) is another outcome variable not related to T, but subject to the same confounders (U) as Y. If it is verified that T has no effect on O, then we can conclude that the estimated effect of T on Y is not biased by U. Formally speaking, $O \perp\!\!\!\perp T|(U,X)$ and $O \not\perp\!\!\!\perp (U,X)$. Similarly, NCE, denoted as E, is another treatment variable that shares the same confounders as T but should not have influence on Y. We have $E \perp\!\!\!\perp Y|(U,X,T)$ and $E \perp\!\!\!\perp O|(U,X,T)$. In this case, the basic assumptions must be revised, and additional assumptions need to be considered.

(1) *Positivity*. If the joint distribution $\mathbb{P}(U,X) > 0$, then the joint conditional density $\mathbb{P}(T,E|U,X) \in (0,1)$. It ensures that all treatment groups could be covered when all characteristics (X and U) are given, supporting the comparison between different groups.
(2) *Latent ignorability* [93]. $Y(t) \perp\!\!\!\perp T|\{X,U\}, \forall t \in T$. It guarantees that the causal effect could be estimated without bias by controlling all confounders (X and U).
(3) *SUTVA*. If treatment $T = t$ and NCE $E = e$, then the outcome and NCO is unique, i.e., $Y = Y(t,e)$ and $O = O(t,e)$. It ensures *consistency* of two exposures (T and E) and *no interference* on two outcomes (Y and O).
(4) *Confounding bridge* [93]. There exists at least one function $b(O,t)$ for all $t \in T$ that could satisfy the condition $\mathbb{E}[Y|U,t] = \mathbb{E}[b(O,t)|U,t]$. This assumption ensures that the effect of the unobservable U on Y can be fully reflected via observing O when controlling T.

With these assumptions mentioned above, we can identify $\mathbb{E}[Y(t)] = \mathbb{E}[b(O,t)], \forall t \in T$. Afterwards, the confounding bridge function can be identified by $\mathbb{E}[Y|E,T] = \mathbb{E}[b(O,T)|E,T]$. Therefore, we can rewrite $ADRF(t) = \int b(O,t)d(O,t)$.

Generative Modeling Methods. Variational Auto-Encoder (VAE) [68] is a popular method for learning representation of proxies. For example, **CEVAE** [83] leverages VAE to learn representation of the hidden confounder V

when estimating the effect of binary treatment. It consists of an inference network and a model network that are all derived from TARNet [122], whose objective is to deduce the nonlinear relationship between X and $V \oplus Y \oplus T$ so as to obtain the approximate solution of $\mathbb{P}(V, X, T, Y)$. The variational lower bound is:

$$\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_{q(V_i|X_i,T_i,Y_i)} [\log p(X_i, T_i|V_i) \ + \log p(Y_i|T_i, Z_i) + \log p(V_i) - \log q(V_i|X_i, T_i, Y_i)], \tag{5}$$

where the first two terms refer to the reconstruction loss and the last two represent the KL divergence. For each sample, it first goes through the inference network to obtain $\mathbb{P}(V|X, Y, T)$. Putting it into the model network gives the value of $\mathbb{P}(Y|T = 1, X)$ and $\mathbb{P}(Y|T = 0, X)$, respectively. ITE is the difference between them. Note that the intention of VAE used here is not to generate samples but to offer better representations of the hidden confounders for the causal estimand.

3.2.2 Instrumental Variable. It is a powerful tool for causal inference with unobserved confounders. A variable Z is regarded as an IV if all the following conditions could be satisfied:

(1) *Relevance.* Z is related to T, i.e., $Z \not\perp\!\!\!\perp T$.
(2) *Exclusion.* Z affects Y only through T, i.e., $Z \perp\!\!\!\perp Y \mid (T, X, U)$.
(3) *Unconfounded instrument.* Z is independent of the confounders X and U, i.e., $Z \perp\!\!\!\perp (X, U)$.

As the unconfounded instrument condition is too strict to be satisfied in real-world applications, the Conditional IV (CIV) [14] is proposed to relax $Z \perp\!\!\!\perp (X, U)$ to $Z \perp\!\!\!\perp U \mid X$. Generally speaking, it is hard to determine suitable IVs or CIVs for the treatments of interest in reality and often requires professional knowledge. Therefore, an IV that meets all the conditions above is called a valid IV, while a weak IV refers to an instrument with weak correlation to the treatment. Even worse, an IV is regarded as an invalid IV with which the aforementioned assumptions will be violated.

3.2.3 Difference-in-Difference Methods. These methods are widely applied when there are repeated observations between different groups over time. A core assumption is needed, which is named *common trends* or *parallel trends*. That is, the average change in Y for the treated group ($T = 1$) would have followed the same trend as the control group ($T = 0$) if no intervention had occurred. In this way, they are powerful tools to tackle time series data of multiple time periods by controlling time-invariant confounders. We denote the time indicator as G, where $G = 0$ means that this sample is recorded before treatment and $G = 1$ refers to the post-treatment condition. Generally, DID [6, 8] defines $Y = \alpha + \beta G + \gamma T + \delta(G \times T) + \varepsilon$, where α, β, γ and δ are the regression coefficients, and ε is an error term. Then, we have:

$$ATT_{DID} = \mathbb{E}[Y|G = 1, T = 1] - \mathbb{E}[Y|G = 0, T = 1] - \mathbb{E}[Y|G = 1, T = 0] + \mathbb{E}[Y|G = 0, T = 0]. \tag{6}$$

Changes-in-Changes (CIC) [8] is an extension of DID that considers unobserved characteristics U. The outcome without treatment is modeled as $Y^N = h(U, G)$, where h is *strictly monotonic* and superscript N means untreated subpopulations with $(T, G) = \{(0, 0), (0, 1), (1, 0)\}$. CIC aims to estimate counterfactual outcome $Y^N_{T,G=1,1}$, abbreviated as Y^N_{11}. Due to the *time invariance assumption* that $U \perp\!\!\!\perp G \mid T$, we can infer how the treated group might have changed without intervention if we know the changes in control group. Specifically, Y_{10} is the factual outcome of treated group before intervention, and quantile $q = F_{Y,00}(Y_{10})$ tells the sample percentage in control group with an outcome $\leq Y_{10}$ before treatment. Then, $F^{-1}_{Y,01}(q)$ finds the outcome $\tilde{y}$ in control group after intervention, satisfying $F_{01}(\tilde{y}) = q$. We regard such $\tilde{y}$ equals Y^N_{11}, and Eq. (6) is rewritten as:

$$ATE_{CIC} = \mathbb{E}[Y^I_{11}] - \mathbb{E}[Y^N_{11}] = \mathbb{E}[Y^I_{11}] - \mathbb{E}[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))], \tag{7}$$

where I refers to $(t, g) = (1, 1)$, and Y^I_{11} is the factual outcome of treated group after receiving T. However, DID and CIC rely on the strong assumptions that do not hold in many practical applications. A common approach to relax these assumptions is to use factor-model based methods, such as synthetic controls [5]. Besides, some

researchers recently introduce a state variable to construct one more difference, and propose a **triple changes estimator [2]**, relaxing those assumptions in another way. Furthermore, DID can also be interpreted as **negative outcome control (NOC) [129]** approach, where pre-exposure outcomes can serve as NCOs. *Distributional equi-confounding assumption* is proposed to support nonparametric identification. Specifically, the mean and variance of Y and NCOs are modeled as functions of X. In this way, the distributions of residuals ϵ_Y and ϵ_O could be flexibly modeled using location-scale models, accommodating different data distributions and confounding structures.

3.2.4 Front-door Criterion. Under the framework of Structural Causal Models (SCM), the front-door criterion addresses unobserved confounding by leveraging observed mediators that fully transmit the causal effect from the treatment to the outcome. Specifically, even when T and Y are confounded, identification is possible if the effect of T on Y operates entirely through a set of mediators M that satisfy the following conditions [39, 103, 104]:

(1) M blocks all the paths from T to Y. Therefore, $\mathbb{P}(Y|\text{do}(T)) = \int_{\mathcal{M}} \mathbb{P}(Y|\text{do}(M))\mathbb{P}(M|\text{do}(T))dM$, where the do-operator means forcing T to take a specific value.
(2) There is no confounding between T and M, and thus $\mathbb{P}(M|\text{do}(T)) = \mathbb{P}(M|T)$.
(3) Conditional on T, there is no confounding between M and Y. Therefore, we have $\mathbb{P}(Y|\text{do}(M)) = \int_{\mathcal{T}} \mathbb{P}(Y|T, M)\mathbb{P}(T)dT$. In summary, we can derive:

$$\mathbb{P}(Y|\text{do}(T)) = \int_{\mathcal{M}} \mathbb{P}(M|T) \left(\int_{\mathcal{T}} \mathbb{P}(Y|T, M)\mathbb{P}(T)dT \right) dM. \tag{8}$$

These probabilities can be directly learned from observational data. This front-door criterion can also be relaxed and extended to estimate indirect causal effect using a doubly robust estimator [33].

3.3 Conclusion and Discussion

When the *unconfoundedness assumption* holds, challenge in this setting is the confounding bias caused by X. **Propensity score** is a widely used method to quantify the influence of X on T. PS has a well-developed theory for identification, but requires explicitly modeling the distribution $\mathbb{P}(T = 1|X)$, which is a strong assumption. Weighting methods based on PS are sensitive to model misspecification and extreme values. Matching and stratification methods have no concerns about extreme values but they require sufficient samples, thus limited in model generalizability. **Doubly robust** methods relax such model correctness assumption, only requiring at least one from the PS or outcome regression model being correctly specified. Instead of relying on these strong assumptions, **covariate balancing** methods directly adjust X through re-weighting in a data-driven manner such as moment constraints. However, they may face the feature selection issue, i.e., treating all variables equally and instead introducing additional noise. In comparison, **tree-based** methods naturally have advantages in feature selection and interpretability. Nevertheless, there is also a risk of over-fitting if the tree grows too deep and the training samples are insufficient. **Representation-based** methods share the same concern about over-fitting, and their interpretability is relatively limited. Nevertheless, they have stronger capabilities of modeling nonlinear relationships especially in complex environments with high-dimensional data. When samples are sufficient, these methods often perform better. Compared to representation learning, **generative models** like GAN have an advantage in capturing the unseen counterfactual distribution. Moreover, they are mainly used in unsupervised learning scenarios without data annotation. However, there may be instability issues during training, including non-convergence, gradient abnormality, and mode collapse.

When considering the existence of unobserved confounders, endogeneity is a key challenge. The typical methods using proxy variables in statistics are called **negative controls**, which have solid theoretical foundations in the identification of treatment effect. However, these methods often require strict assumptions that are hard to satisfy in practical applications. In computer science, **generative models** like VAE are applied to recover latent

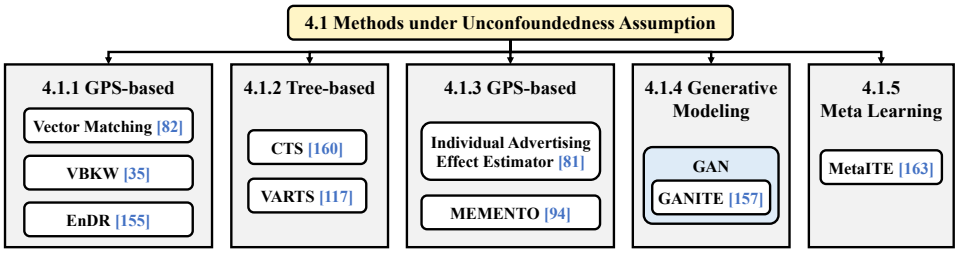


Fig. 4. Categorization of multi-valued treatment methods without unobserved confounders.

confounders from observed data for counterfactual prediction. Proxies are supposed to be sufficient to cover all the unmeasured confounders, making such recovery process a challenging task. **Instrumental variables** are auxiliary information to help estimate the influence of T on Y, but how to provide valid IVs is also a hard problem that often requires expert knowledge. **DID** and **CIC** methods are particularly suitable for evaluating the effects of non-random treatment allocation, such as policy change and law implementation. But their validity relies on the *parallel trends assumption* that is difficult to verify in practice. They are also not good at solving situations with heterogeneous treatment effects.

4 Multi-valued Treatment

In this section, we introduce relevant methods that estimate the causal effect of multi-valued treatment in two cases. The first case is that the *unconfoundedness assumption* holds, and the second case is that there exist unobserved confounders.

4.1 Under Unconfoundedness

Existing works for multi-valued treatment under the *unconfoundedness assumption* are organized in Fig. 4. They can be further divided into 5 categories: GPS-based methods, tree-based models, representation-based methods, generative modeling methods, and meta learning methods.

4.1.1 GPS-based Methods. The **Generalized Propensity Score (GPS)** [55] is an extension derived from PS, whose definition is given in Eq. (1). GPS is proposed as a solution for the settings of multiple treatments with discrete values and its expression is given as below:

$$e(T, X) = f_{T|X}(T|X), \tag{9}$$

where $f_{T|X}(T|X)$ means the conditional density of T given X. Suppose there are m treatments, then the GPS can be rewritten in the form of a vector as $\boldsymbol{e}(X) = [e(t_1, X), \ldots, e(t_m, X)]$. Afterwards, many approaches using PS in the binary treatment setting, such as those introduced in Section 3.1.1, can also be extended to the multi-valued treatment setting by using GPS.

Vector Matching [82] aligns subjects with similar $e(X)$, usually applying a multinomial regression model to determine a common support region for multiple treatments. For each $t \in \mathcal{T}$,

$$\begin{aligned} e(t, X)^{(low)} &= \max\left(\min\left(e(t, X|T = t_1)\right), \ldots, \min\left(e(t, X|T = t_m)\right)\right) \\ e(t, X)^{(high)} &= \min\left(\max\left(e(t, X|T = t_1)\right), \ldots, \max\left(e(t, X|T = t_m)\right)\right) \end{aligned} \tag{10}$$

where $e(t, X|T = l)$ refers to the treatment assignment probability for t among those subjects that received treatment l. The purpose is to maximize the similarity between matched sets while minimizing the bias between covariate distributions. Subjects with $e(t, X) \notin (e(t, X)^{(low)}, e(t, X)^{(high)})\ \forall t \in \mathcal{T}$ will be discarded, followed by re-fitting the GPS model. Afterwards, k-means clustering [44] is applied to divide all the subjects into several

clusters, where those within the same cluster are similar on one or more GPS components. It is guaranteed that there is at least one subject of each treatment in each cluster. A pair of subjects will be matched if they belong to the same subclass.

Vector-based Kernel Weighting (VBKW) [35] is a hybrid method that combines kernel weighting and vector matching. Let l denote the samples of treatment t, and j denote the samples of treatment $t' \neq t$. Based on propensity score vectors $\boldsymbol{e}(X) = [e(t_1, X), \ldots, e(t_m, X)]$, $l_{matched} \subset l$ includes samples in l that are successfully matched to one or more samples in j. The weight for ATT of treatment t is $w_{ATT,i} = \begin{cases} 1 & , \forall i \in l_{matched} \\ k_i(D_{lt}), & \forall i \in j \end{cases}$, where $k_i(\cdot)$ refers to the Epanechnikov kernel, and $D_{lt} = |\hat{e}(X_l, t) - \hat{e}(X_j, t)|$. Weight $w_{i,ATT'}$ for treatment t' can be constructed in a similar manner. Weights for estimating ATE of t vs. t' can be expressed as $w_{i,ATE} = w_{i,ATT} + w_{i,ATT'}$. Finally,

$$ATE_{t,t'} = \frac{\sum_{i \in l} Y_i d_i(t) w_{i,ATE}}{\sum_{i \in l} d_i(t) w_{i,ATE}} - \frac{\sum_{i \in j} Y_i d_i(t') w_{i,ATE}}{\sum_{i \in j} d_i(t') w_{i,ATE}}, \tag{11}$$

where $d_i(t) = \mathbb{I}(T_i = t)$ is an indicator variable.

Ensemble Doubly Robust (EnDR) [155] follows the doubly robust method descried in Eq. (3), but the estimator $e(\cdot)$ is substituted by Eq. (9). EnDR establishes the propensity score estimator $e(\cdot)$ and outcome regressors $\mu_t(\cdot)$ in an ensemble manner, i.e., choosing the top 1 or using the weighted average from multiple models. Specifically, it uses rank aggregation strategy to ensemble linear regression, CBPS (estimating Y only), random forest, and generalized boosted model.

4.1.2 Tree-based Methods. Decision makers are interested about casting which campaign (multi-valued treatment) could obtain the best uplift (ITE or CATE). It can be seen as a map function, i.e., $h(\cdot) : \mathbb{X}^d \rightarrow \{1, \ldots, m\}$. The goal is to figure out the optimal treatment with the best expected response by $h^*(x) \in \arg\max_{t=1,\ldots,m} \mathbb{E}[Y|X = x, T = t]$. Tree structure is naturally suitable for this.

Contextual Treatment Selection (CTS) [160] divides the whole feature space into disjoint subspaces, where each subspace corresponds to a specific treatment. These subspaces are represented as the leaf nodes of the decision tree. Each leaf provides the probability of a sample falling into that subspace (i.e., adopting a specific treatment) and the expected response (potential outcome). During the construction of each tree, a recursive binary splitting approach is applied and the goal is to maximize the expected response. Splitting criterion is to measure the increase in the holistic expected response $\Delta\mu(s)$ of a candidate split s that divides a leaf node ϕ into ϕ_l and ϕ_r:

$$\begin{aligned} \Delta\mu(s) = {} & \mathbb{P}(X \in \phi_l | X \in \phi) \max_{t_l=1,\ldots,m} \mathbb{E}[Y|X \in \phi_l, T = t_l] \\ & + \mathbb{P}(X \in \phi_r | X \in \phi) \max_{t_r=1,\ldots,m} \mathbb{E}[Y|X \in \phi_r, T = t_r] - \max_{t=1,\ldots,m} \mathbb{E}[Y|X \in \phi, T = t]. \end{aligned} \tag{12}$$

In details, $\mathbb{P}(X \in \phi' | X \in \phi)$ can be regarded as the probability of a subject further falling into ϕ' conditioned on already divided to ϕ. It can be rewritten as $\hat{p}(\phi'|\phi) = \sum_{i=1}^{N} \mathbb{I}\{X_i \in \phi'\} / \sum_{i=1}^{N} \mathbb{I}\{X_i \in \phi\}$. Let $\hat{y}_t(\phi')$ denotes the expected response of subspace ϕ' given treatment t, which is defined as:

$$\hat{y}_t(\phi') = \begin{cases} \hat{y}_t(\phi) & , \text{ if } n_t(\phi') < \text{min\_split} \\ \dfrac{\left(\sum_{i=1}^{n} y_i \mathbb{I}\{X_i \in \phi'\} \mathbb{I}\{T_i = t\} + \hat{y}_t(\phi) \cdot \text{n\_reg}\right)}{\left(\sum_{i=1}^{n} \mathbb{I}\{X_i \in \phi'\} \mathbb{I}\{T_i = t\} + \text{n\_reg}\right)}, & \text{otherwise} \end{cases} \tag{13}$$

where min_split is a user-defined parameter, meaning the minimum number of samples required to perform a split. Another parameter n_reg, usually a small positive integer, is provided as a regularity term to avoid misleading from outliers. The response increase can be expressed as:

$$\Delta\hat{\mu}(s) = \hat{p}(\phi_l|\phi) \times \max_{t=1,\ldots,m} \hat{y}_t(\phi_l) + \hat{p}(\phi_r|\phi) \times \max_{t=1,\ldots,m} \hat{y}_t(\phi_r) - \max_{t=1,\ldots,m} \hat{y}_t(\phi). \tag{14}$$

Construction of a tree is completed when there is no split to conduct, or the samples in the node have the same response value. CTS also creates a forest to alleviate over-fitting of a single tree.

VAriance Reduced Treatment Selection (VARTS) [117] demonstrates that CTS requires a large amount of training data, and the estimator maximized in CTS is biased against the true metric. Therefore, it proposes a variance reduced estimator based on the doubly robust estimation. Specifically, the expected response is:

$$\hat{V}_{varts}(\phi, t) = \frac{1}{n_\phi} \sum_{i:x_i \in \phi} \left(\frac{\left(Y_i^{obs} - \hat{\mu}_i^{(t)}\right) \mathbb{I}\{T_i = t\}}{p^{(t)}} + \hat{\mu}_i^{(t)} \right). \tag{15}$$

Accordingly, the split criterion can be described as:

$$\hat{s} = \arg\max_{s \in \mathcal{S}} \hat{p}\left(\phi_l(s)|\phi\right) \times \max_{t_l \in \mathcal{T}} \hat{V}_{varts}\left(\phi_l(s), t_l\right) + \hat{p}\left(\phi_r(s)|\phi\right) \times \max_{t_r \in \mathcal{T}} \hat{V}_{varts}\left(\phi_r(s), t_r\right). \tag{16}$$

4.1.3 Representation-based Methods. The CFR framework is extended to solve the problem of multi-valued treatment as well. The most crucial modification lies in how to balance the covariate distributions across multiple groups (corresponding to multiple discrete T values), and how to model the hypothesis function(s) applicable to all these groups.

An intuitive way to control the confounding bias is using IPM to constrain all the possible pairs of different treatment groups, with a total number of C_m^2 for m treatment values. **Individual Advertising Effect Estimator [81]** proposes a *transitivity assumption* so that only the IPM between adjacent treatment pairs needs taken into account. All groups with various treatment assignments share the same hypothesis network denoted as $h(\Phi(X), T)$. The objective function is:

$$\begin{aligned} \min_{h,\Phi} \frac{2}{n} \sum_{i=1}^{n} w_{T_i} \cdot \mathcal{L}\left(h(\Phi(X_i), T_i), Y_i\right) + \lambda \cdot \mathcal{R}(h) + \beta \cdot \sum_{j=1}^{m-1} \mathrm{IPM}_{\mathrm{G}}(p_\Phi^{T=T_j}, p_\Phi^{T=T_{j+1}}) \\ - w_{T_1} \sum_{i=1}^{n} \mathcal{L}\left(h(\Phi(X_i), T_i), Y_i\right) \mathbf{1}_{T_i=T_1} - w_{T_m} \sum_{i=1}^{n} \mathcal{L}\left(h(\Phi(X_i), T_i), Y_i\right) \mathbf{1}_{T_i=T_m} \end{aligned} \tag{17}$$

where w_{T_i} is the proportion of units applying T_i, and $\mathcal{R}$ is a model complexity term.

MEMENTO [94] follows the intuitive idea that using C_m^2 pairs of Maximum Mean Discrepancy (MMD) constraints between all treatment groups to address the confounding bias, and constructing m hypothesis functions h_{t_i} for counterfactual prediction. Its key contribution lies in introducing the Expected Precision in Estimation of Heterogeneous Effect (PEHE) loss [49] to the setting of multi-valued treatment. Since the counterfactual loss can not be directly calculated by observational data, it deduces an upper bound of $\mathcal{L}_F(h) + \mathcal{L}_{CF}(h)$. Accordingly, the goal is to minimize:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mu_{T_i}} \mathcal{L}(Y_i, h_{T_i}(\Phi(X_i))) + \mathcal{R}(\Phi, h_{1,\ldots,m}) + \alpha \cdot \sum_{t \neq t'} \mathrm{MMD}(\mathbb{P}(\Phi(X)|t), \mathbb{P}(\Phi(X)|t')). \tag{18}$$

4.1.4 Generative Modeling Methods. Although **GANITE [157]** is introduced under the setting of binary treatment, it can be easily developed to tackle multi-valued treatment as well. The key modification lies in the discriminator of the counterfactual block, whose goal is to determine which value of the treatments correspond to the real outcome in Y_i^F. Other parts of GANITE remains unchanged. In this way, it can estimate the potential outcome of different treatment values.

4.1.5 Meta Learning Methods. There is an issue that the sample sizes in different groups are often imbalanced. **MetaITE [163]** introduces the concept of meta-learning, regarding a treatment group with sufficient samples as a source domain that is firstly used to train the base model. Conversely, it treats a group with few samples as a target domain, on which the base model is fine-tuned. Two core components consist of the base model, including a feature extractor to obtain balanced embeddings across multiple domains, and an inference network to estimate

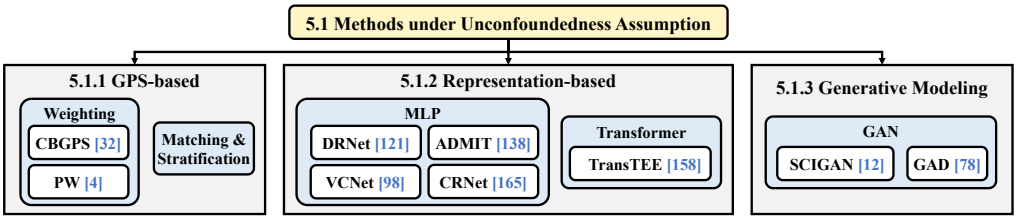


Fig. 5. Categorization of continuous treatment methods without unobserved confounders.

potential outcomes. Objective function comprehensively considers the prediction errors on two kinds of domains, together with the MMD metric constrained on the their embeddings for domain balance.

4.2 Conclusion and Discussion

Many methods addressing the multi-valued treatment discussed in this section are extended from those initially proposed for binary treatments. These extensions involve a fundamental shift from binary to multi-class classification. For propensity score-based methods, this means extending Eq. (1) to Eq. (9), supporting the estimation of propensity scores for m treatment categories (e.g., via multinomial logistic regression). Tree-based methods, which typically determine the best treatment value from a binary set $\{0, 1\}$, are adapted to select from $\{1, \ldots, m\}$ by modifying the splitting criteria or employing multi-output tree structures. Similarly, GANITE extends its discriminator's role from identifying 2 labels to distinguishing between m distinct treatment labels.

Another key extension is the balancing between multiple treatment groups, in terms of covariate distributions or representations. Binary treatment can be regarded as a special case of multiple treatments. Given that the set of all possible treatments is finite (multi-valued and bundle treatments), this can be naturally achieved by incorporating treatment-specific representation networks, hypothesis networks, and MMD constraints. Methods designed for infinite treatment sets (continuous treatment) will be further discussed in Section 5.

Finally, some methods demonstrate generality. As stated in Section 3.2, methods such as proxy and instrumental variables do not restrict the type of T in their solutions. Therefore, they can be naturally applied to continuous or bundle settings, already covering the cases of multiple discrete values. We will discuss these methods in Section 5.2.1 and Section 5.2.2.

5 Continuous Treatment

Another setting included in the complex treatment is that the value of treatment could be continuous. Relevant methods will be introduced from two aspects, covering the methods obeying *unconfoundedness assumption* and those considering the unobserved confounders.

5.1 Under Unconfoundedness

As shown in Fig. 5, we will introduce four kinds of methods in this section, including GPS-based methods, doubly robust methods, representation-based methods, and generative modeling methods.

5.1.1 GPS-based Methods. The GPS mentioned in Eq. (9) is designed to quantify the possibility of a unit receiving a certain treatment from multiple choices. Depending on the *smoothness assumption* in Section 2, it can be naturally extended for the estimation of continuous treatment effect. Researchers often explicitly model GPS of continuous treatments by Gaussian distribution [16, 50, 54, 111].

Weighting Methods. Inspired by **IPW** [85], the **inverse of generalized propensity scoring (IGPS)** [111] re-weights samples using the reciprocal of GPS. To address the issue of extreme values in the denominator, researchers propose **Stabilized IGPS (SIGPS)** [111]. Formally, $w_i^{SIGPS} = \frac{f(T_i)}{e(T_i, X_i)}$, and $f(T_i)$ is the probability density of

treatment T_i. However, some researchers point out that these methods are sensitive to model misspecification [128, 167]. Therefore, optimal balancing weighting is studied, which is concentrated on achieving a direct balance in treatment assignment without the explicit specification of the conditional density $f(T|X)$. These methods theoretically have the doubly robust property [159]. **Covariate Balancing Generalized Propensity Score (CBGPS) [32]** is an extension of CBPS [53], which is mentioned in Section 3, to the setting of continuous treatment. It aims to eliminate the correlation between treatment T and covariates X. To ensure the balancing property of the GPS, CBGPS formulates the moment conditions as $\mathbb{E}[w^{CB}TX] = \mathbb{E}[T]\mathbb{E}[X]$, where the wights w^{CB} are constrained by $\mathbb{E}[w^{CB}] = 1$. To maximize its empirical likelihood, the objective can be formally defined as $\arg\min_{w^{CB}\in\mathcal{W}} \sum_{i=1}^{n} \log w_i^{CB}$. **Permutation Weighting (PW) [4]** is another balancing weighting method whose main idea is to randomly conduct permutation on T of the original data $\mathcal{P}$ so as to satisfy the balance condition. That is to say, the assignment of T is independent of X in the permuted data $\mathcal{Q}$. Afterwards, a classifier is trained to tell whether a sample is from $\mathcal{P}$ (labeld as $C = 1$) or $\mathcal{Q}$ (labeled as $C = 0$). Once it is trained, the sample weight can be measured through probability density ratio, i.e., $w^{PW} = \frac{\mathbb{P}(C=1|T_i,X_i)}{\mathbb{P}(C=0|T_i,X_i)}$.

Matching and Stratification Methods[1]**.** The matching method for binary treatment is also generalized to the continuous treatment setting [84]. For example, propensity function [54] is proposed to offer a balancing function, not a one-dimensional score, compared to the GPS. Suppose there is a unique propensity function $\theta(t, x)$ for all $t \in \mathcal{T}, x \in \mathcal{X}$, such that $e(t, x)$ depends on x through $\theta(t, x)$. The stratification principle based on the propensity function can be expressed as:

$$ADRF(t) = \int f\big(Y(t)|T = t, \theta(t, x)\big) f(\theta(t, x)) d\theta(t, x), \tag{19}$$

where $f(\cdot)$ refers to the probability density. Researchers also point out that the fundamental concept underlying current matching or stratification methods is discretization. Effectiveness of these methods mainly depends on the choice for distance metrics and the number of strata [54]. Advantage of these methods lies in their strong interpretability and they have no concerns about the issue of extreme values. However, they usually require sufficient samples for data fitting, and need discretization of the continuous treatment.

5.1.2 Representation-based Methods. Backbones of the representation learning to facilitate causal inference often include Multi-Layer Perceptron (MLP) and transformer [137].

MLP methods. They mainly follows the basic architecture of CFR in Section 3.1.5, but need to make additional improvements in discretizing treatment T. **DRNet [121]** handles this problem by utilizing a bucket method to assign continuous variables into discrete intervals. Each interval corresponds to a prediction head. However, it may disrupt the continuity of causal effects and lead to jump discontinuity at the boundaries of buckets. Therefore, **VCNet [98]** proposes varying coefficient prediction heads to retain the continuity of dose response curves. Specifically, the prediction head can be described as $\mu^{NN}(t, x) = \mathbb{E}[Y|T = t, X = x] = f_{\theta(t)}(p)$, where $f_{\theta(t)}$ is a neural network with variable parameters $\theta(t) = [\theta_1(t), \ldots, \theta_{d_\theta}(t)]^\top \in \mathbb{R}^{d_{\theta(t)}}$, $d_{\theta(t)}$ is the dimension of $\theta(t)$, and p is the feature extracted from the conditional density estimator $\pi(t|x)$. Formal definition of each dimension is $\theta_i(t) = \sum_{l=1}^{L} a_{i,l}\varphi_l^{NN}(t)$, where $\{\varphi_l^{NN}\}$ are the spline basis and $a_{i,l}$ are the coefficients. **ADMIT [138]** is proposed to theoretically support the alleviation of selection bias in the setting of continuous treatments. Besides the balanced representation $\Phi(X)$ and predicted potential outcome $h(T, \Phi(X))$ from varying coefficient models, it also learns weights $w(T, \Phi(X))$ via importance sampling. Such w are utilized in weighted prediction loss $w(T, \Phi(X)) \cdot \mathcal{L}(h(T, \Phi(X)), Y)$, together with IPM constraints between the factual and counterfactual distributions $\mathrm{IPM}_G(w \cdot \mathbb{P}(X), \mathbb{P}(X \mid T))$. Note that many representation-based methods learn the representation $\Phi(X)$ by constraining $T \perp\!\!\!\perp X \mid \Phi(X)$, thus neglecting much of the useful information in X. **CRNet [165]** theoretically

[1]In the context of continuous treatment setting, the stratification method can be regarded as a specific instance of the non-bipartite matching method.

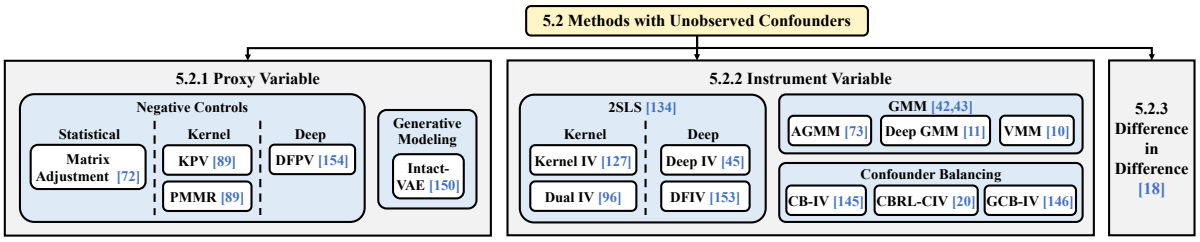


Fig. 6. Categorization of continuous treatment methods with unobserved confounders.

demonstrates the importance of the balancing and prognostic representations, i,e. $T, Y \perp\!\!\!\perp X \mid \Phi(X)$. It learns a double balancing representation with such constraints via contrastive learning so as to estimate the HDRF while preserving the continuity of treatments.

Transformer methods. Compared to MLP, Transformer does better in capturing the internal interactions within X, as well as the interaction between T and X. For example, **TransTEE [158]** attempts to introduce transformer [137] into the context of causal inference for continuous treatment. It consists of three key components: embedding layers to map X and T into a low-dimensional space, self-attention layers to capture the internal interactions between the two embeddings, and cross-attention layers to enables the embedding of T focused on X that are most relevant to Y. Afterwards, the objective of TransTEE is to minimize the MSE of outcome regression.

5.1.3 Generative Modeling Methods. Inspired by GANITE [157], **SCIGAN [12]** estimate the IDRF for continuous treatments by employing a generator that directly produces response curves. A core feature of SCIGAN is its hierarchical discriminator architecture. The first level involves a treatment discriminator, tasked with distinguishing between discrete types of treatments. Subsequently, within the dosage interval corresponding to each treatment type, a second-level discriminator, the dosage discriminator, focuses on identifying the factual (continuous) dosage. This hierarchical design allows for a more nuanced estimation of causal effects by effectively separating the influences of treatment type and dosage. The generative adversarial aspect is crucial: the generator creates counterfactual outcomes, aiming to fool the discriminator into believing they are factual, while the discriminator learns to differentiate between factual and generated outcomes. This adversarial training process encourages the generator to learn the true distribution of unobserved counterfactuals, enabling the estimation of personalized treatment effects for continuous interventions. Following this, **Generative Adversarial De-confounding (GAD) [78]** adopts GANs with permutation weighting [4] to learn the GPS for ADRF. Specifically, in GAD, the generator is used to learn the distribution of GPS, while the discriminator strives to distinguish whether the data are from the permuted distribution or the fitted distribution. After training, the weight $\hat{w}(X,T)$ output from the generator can be regarded as IGPS, and $\hat{ADRF} = \sum_{i:T_i=t} \hat{w}_i \cdot Y_i(t)$.

5.2 With Unobserved Confounders

In the presence of unobserved confounders, whether in binary or continuous treatment cases, one of the most promising methods is the use of proxy variables. With the advancement of big data collection and the improvement of various data policies, even if some variables cannot be directly observed, we can capture their effects by leveraging their relationships with certain observed variables. This approach enables us to mitigate the bias caused by unobserved confounders, thereby obtaining more accurate causal estimates. Note that instrumental variables (IVs) are a special type of proxy variable. For exogenous instrumental variables, the fields of statistics and econometrics have developed numerous IV regression algorithms to study treatment effects. Therefore, we will further discuss these proxy variable methods and IV methods as shown in Fig. 6.

5.2.1 Proxy Variable. Proxy variables are used as substitutes for unobserved confounders. By leveraging variables that are correlated with the unobserved confounders, proxy variable methods aim to control for hidden biases and

enable more accurate causal inference. The use of proxy variables is particularly valuable in complex datasets with high-dimensional or missing information, as they allow researchers to approximate the influence of unobserved factors through related, observable data. As big data sources expand and data policies improve, proxy variable methods continue to gain traction in both binary and continuous treatment scenarios, providing an effective means of adjusting for hidden confounding. Extensive research on proxy variables encompasses both negative control methods and generative modeling methods.

Negative Controls have been briefly introduced in Section 3.2.1. We classify the relevant methods into statistical, kernel, and deep methods. **(1) Statistical Methods.** We introduce the **Matrix Adjustment Method [72]**, a probability distribution-based correction technique to adjust for the impact of U with O and E. The core idea is utilize relationship $\mathbb{P}(U|T,Y) = M(O,U)^{-1}\mathbb{P}(O|T,Y)$ to recover $\mathbb{P}(T,Y,U)$ from $\mathbb{P}(T,Y,O)$ and $\mathbb{P}(O|U)$ by adjusting $M(O,U)$, where $M(O,U)$ is the matrix of conditional probabilities between O and U. With linear structural equation model, we have $ATE_{yt} = \frac{\sigma_{ty} - \alpha_o u^2 \sigma_{tu}\sigma_{yu}}{\sigma_{tt} - \alpha_o u^2 \sigma_{tu}^2}$, where $\alpha_o u$ is the regression coefficient of O to U and σ is the covariance.

(2) Kernel Methods, especially those utilizing Reproducing Kernel Hilbert Spaces (RKHS) [102], are able to capture complex nonlinear relationships by mapping data to high or infinite dimensional spaces. They are also widely used in causal effect estimation, and we give two examples as below.

- **Kernel Proxy Variable (KPV) [89].** It is a two-stage regression method. In the first stage, KPV estimates the relationship between potential confounders U and proxy variables O by learning conditional mean embedding $\mu_O|T,X,E$ via kernel regression. In the second stage, estimation function is learned as a mapping $\eta[\cdot]$ from $\mu_O|T,X,E$ to Y by minimizing:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \eta\left[\Phi(T_i,X_i)\otimes\hat{\mu}_O|T_i,X_i,E_i\right]\right)^2 + \lambda\|\eta\|^2_{H_{TXO}}, \tag{20}$$

  where n is the total number of samples in the dataset, $\Phi(\cdot)$ means a kernel feature mapping from the input space into the RKHS H_{TXO}, operator $\otimes$ refers to the tensor product, and λ is a hyper-parameter for regularization.
- **Proxy Maximum Moment Restriction (PMMR) [89].** It is a single-stage approach aimed at finding the structural function $h(T,X,O)$, best satisfying the Conditional Moment Restriction (CMR) so that $\mathbb{E}[Y - h(T,X,O)|T,X,E] = 0$ almost certainly holds. Kernel methods here addressing the CMR problem is to find a closed-form solution in the kernel space.

(3) Deep Methods. Deep learning models can also serve as powerful tools for treatment effect estimation, especially in cases with high-dimensional data and complex relationships. We introduce **Deep Feature Proxy Variable (DFPV) [154]** here for instance. Its main idea is similar to that of KPV, but it uses deep neural networks rather than kernel functions when learning the feature mapping. Different neural nets are trained to learn treatment features in different stages, denoted as $\phi_{\theta_{T(1)}}$ and $\psi_{\theta_{T(2)}}$ respectively. Objective in the first stage is to learn weights $\boldsymbol{V}$ and parameters $\{\theta_{T(1)},\theta_E\}$ by minimizing the following empirical loss:

$$\hat{\mathcal{L}}_1\left(\boldsymbol{V},\theta_{T(1)},\theta_O\right) = \frac{1}{n}\sum_{i=1}^{n}\left\|\psi_{\theta_O}(O_i) - \boldsymbol{V}\left(\phi_{\theta_{T(1)}}(T_i)\otimes\phi_{\theta_E}(E_i)\right)\right\|^2 + \lambda_1\|\boldsymbol{V}\|^2, \tag{21}$$

Fixing $\{\boldsymbol{V},\theta_{T(1)},\theta_E\}$, the second stage is to learn $\{\boldsymbol{u},\theta_{T(2)},\theta_O\}$ by minimizing:

$$\hat{\mathcal{L}}_2\left(\boldsymbol{u},\theta_E,\theta_{T(2)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{u}^\top\left(\psi_{\theta_{T(2)}}(T_i)\otimes\hat{\boldsymbol{V}}\left(\phi_{\hat{\theta}_{T(1)}}(T_i)\otimes\phi_{\hat{\theta}_E}(E_i)\right)\right)\right)^2 + \lambda_2\|\boldsymbol{u}\|^2. \tag{22}$$

Generative Modeling Methods. Inspired by CEVAE, **Intact-VAE [150]** proposes a new variational autoencoder variant for estimating treatment effects under unobserved confounding via the prognostic score. Although

various methods for causal inference under unobserved confounding have been proposed, these methods are often limited to restrictive parametric or linear models , and some lack practical estimation methods or a complete theoretical framework for general non-linear scenarios. Therefore, the identification and estimation of treatment effects using proxy variables via deep latent variable methods remain a significant challenge requiring further exploration.

5.2.2 Instrumental Variable (IV). An instrumental variable (IV) is an exogenous variable that is not affected by unobserved confounders. It could be regarded as a special case of "negative controls" or "proxies" for unmeasured confounders in advanced causal inference methods [93]. It helps address unmeasured confounding by providing an exogenous source of variation for the exposure. This "cleans" the exposure's effect, so the residuals in the final regression ideally reflect only random error, rather than systematic bias from unmeasured confounders. Furthermore, IV methods are naturally designed to address complex treatments and outcomes, making them widely used in statistics and econometrics. For a comprehensive overview of instrumental variable applications in causal inference, refer to the detailed survey [147].

Two-Stage Least Square (2SLS) [134]. Although Ordinary Least Square (OLS) [144], a direct regression method, is widely used for causal effect estimation, the causality obtained is distorted when unobserved confounders exist. Instead, using two-stage regressions combined with IVs can achieve good performance. We denote the IVs as Z, and their properties have been introduced in Section 3.2.2. Most IV methods assume that the data generation process follows additive noise models. Therefore, we model it as the following models and exclude X from the model for clarity, i.e., $T = f(Z) + \epsilon_T, Y = g(T) + \epsilon_Y, Z^\top \epsilon_T = Z^\top \epsilon_Y = 0$. Assuming that f and g are linear, we can apply 2SLS to remove the confounding bias introduced by unobserved confounders. The details of the two-stage procedure are outlined as follows.

- Treatment Regression Stage. Estimator α regresses T from Z:

$$\hat{\alpha} = (Z^\top Z)^{-1} Z^\top T = (Z^\top Z)^{-1} Z^\top (Z\alpha + \epsilon_T) = \alpha, \tag{23}$$

$$\hat{T} = \mathbb{E}[T|Z] = Z\alpha. \tag{24}$$

- Outcome Regression Stage. Estimator β regresses Y given $\hat{T}$ derived from the first stage:

$$\hat{\beta}_{\text{2SLS}} = (\hat{T}^\top \hat{T})^{-1} \hat{T}^\top Y = (\hat{T}^\top \hat{T})^{-1} \hat{T}^\top (\hat{T}\beta + \epsilon_Y) = \beta, \tag{25}$$

$$\hat{Y} = \mathbb{E}[Y|\hat{T}] = \hat{T}\beta. \tag{26}$$

In more common real-world scenarios where f and g are nonlinear, numerous methods within this framework employ different machine learning techniques to learn the complex nonlinear relationships in the models, thus extending the traditional IV approach to nonlinear scenarios. **(1) Kernel Methods.** RKHS can be used here to capture the nonlinear causal estimands. We will introduce the following two methods.

- **Kernel IV [127].** We define two measurable positive definite kernels, including $k_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Therefore, we have two basis functions, $\psi : \mathcal{T} \to \mathcal{H}_{\mathcal{T}}, t \mapsto k_{\mathcal{T}}(t, \cdot)$ and $\phi : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}}, z \mapsto k_{\mathcal{Z}}(z, \cdot)$. Structural function $\mu(\cdot)$ in the first stage is:

$$\mu(Z) = e(\phi(Z)) = \mathbb{E}[\psi(T)|Z]. \tag{27}$$

  The objective is to optimize $e \in E : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{T}}$ by kernel ridge regression. As for the second stage, the structural function $g(\cdot)$ is formally defined as:

$$g(T) = h\left(\psi(T)\right) = \mathbb{E}[Y|\hat{\psi}(T)]. \tag{28}$$

  The objective is to optimize $h \in H : \mathcal{H}_{\mathcal{T}} \to \mathcal{H}_{\mathcal{Y}}$. When X is included in the model, Kernel IV can be performed by simply combining X with T and X with Z individually.

- **Dual IV [96].** The outcome structural function $g(T)$ can be rewritten as:

$$\mathbb{E}[Y|Z,X] = \mathbb{E}[g(T,X)|Z,X] + \mathbb{E}[\epsilon_Y|X] = \int g(T,X)\, dF(T|Z,X), \tag{29}$$

  where $dF(T|Z,X)$ is the conditional treatment distribution obtained from the treatment regression. By denoting $W = Y \oplus Z \oplus X$ and $\tilde{T} = T \oplus X$, dual IV reformulates the two-stage IV-based regression as a convex-concave saddle-point problem:

$$\min_{g\in\mathcal{G}} \max_{u\in\mathcal{U}} \mathbb{E}_{TW}[g(\tilde{T})u(W)] - \mathbb{E}_W[l^*(Y,u(W))], \tag{30}$$

  where $\mathcal{U}(\Omega) = \{u(\cdot) : \Omega \to \mathbb{R}\}$ is the entire space of functions and Ω is the support of W.

Deep neural networks (DNNs) are also applied to the two-stage regression, called **(2) Deep Methods** here. These methods often make weaker assumptions about the data generation process.

- **DeepIV [45].** Inspired by Eq. (29), the counterfactual prediction can be described as an optimization problem by minimizing the following objective:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \int_T g(T,X_i)\, dF(T|Z_i,X_i)\right)^2. \tag{31}$$

  In the first stage, term $F(T|Z,X)$ can be quantified by a deep neural network $\pi_\phi(Z,X)$ with loss $l(T,\pi_\phi(Z,X))$. In the second stage, another network h_θ is learned to approximate the potential outcome $g(T,X)$. Therefore, objective of DeepIV can be rewritten as minimizing:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \int_T h_\theta(T,X_i)\, d\hat{F}_\phi(T|Z_i,X_i)\right)^2. \tag{32}$$

- **Deep Feature Instrumental Variable (DFIV) [153].** Its main idea is to use DNNs to learn adaptive deep features as kernel basis in the 2SLS approach. Specifically, three DNNs $\{f_\phi, g_\xi, u_\psi\}$ are established to learn the feature mapping of $\{Z,X,T\}$, respectively. Tasks of the two stages are to minimize the following objectives:

$$\frac{1}{n}\sum_{i=1}^{n}\|u_\psi(T_i) - \boldsymbol{A}f_\phi(Z_i)\otimes g_\xi(X_i)\|^2 + \lambda_1\|\boldsymbol{A}\|^2, \text{ and } \frac{1}{n}\sum_{i=1}^{n}\|Y_i - \boldsymbol{B}u_\psi(T_i)\otimes g_\xi(X_i)\|^2 + \lambda_2\|\boldsymbol{B}\|^2, \tag{33}$$

  where $\boldsymbol{A}$ and $\boldsymbol{B}$ refer to the matrix parameters learned in each stage, and λ_1 and λ_2 are hyper-parameters for regularization in case of over-fitting.

Generalized Method of Moments (GMM) [42, 43]. Although previous works perform well in estimating the expected value, their standard errors are inconsistent with the ground truth. Therefore, conditional moment restrictions can be used when facing heteroskedasticity of unknown form. Specifically, d^Z instruments give a set of d^Z moments, i.e., $l_i(g) = z_i^\top u_i = z_i^\top(y_i - g(t_i,x_i))$ for $i = 1,\ldots,n$. The intuition of GMM is to choose a suitable estimator for g so that $\mathbb{E}[l(g)] = 0$. Recently, many researchers reformulate these conditional moment problems as a minimax optimization problem [29]. Taking **Adversarial GMM (AGMM) [73]** as an example, its objective is:

$$\arg\inf_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}\left[(Y - h_\phi(T,X))f_\psi(Z,X)\right] - \lambda_1\|\psi\|^2 - \mathbb{E}\left[f_\psi^2(Z,X)\right] + \lambda_2\|\phi\|^2, \tag{34}$$

where h is a network aimed at setting moments as close to 0 as possible, while f is another adversary network to identify moments that are violated for the chosen h. When there are infinite moment conditions, **Deep GMM [11]** is to construct an optimal combination of moment conditions:

$$\arg\inf_{\phi\in\Phi}\sup_{\psi\in\Psi}\mathbb{E}\left[(Y - h_\phi(T,X))f_\psi(Z,X)\right] - \frac{1}{4}\mathbb{E}\left[(Y - h_\phi(T,X))^2 f_\psi^2(Z,X)\right]. \tag{35}$$

There is also a **Variational Method of Moments (VMM) [10]**, which transforms the original GMM problem into a minimization one that includes an inner maximization problem:

$$\arg\min_{\theta\in\Theta}\sup_{f\in\mathcal{F}}\mathbb{E}\left[f(Z)^\top\rho(X;\theta)\right]-\frac{1}{4}\mathbb{E}\left[\left(f(Z)^\top\rho(X;\hat{\theta})\right)^2\right]-\|f\|^2. \tag{36}$$

Confounder Balancing Methods. When there are interactions among variables, IVs will have limited effect on T, called weak IVs. Therefore, some researchers consider the joint effect of X and Z. **Confounder Balanced IV Regression (CB-IV) [145]** models a more general causal relationship that $T = f_1(Z,X) + f_2(X,U)$, $Y = g_1(T,X) + g_2(T)g_3(U) + g_4(X,U)$ and $Z \perp\!\!\!\perp U, X$. In the first stage, it regresses T with Z and X by minimizing $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(T_i - T_i^j)^2$ and $\hat{T}_i^j \sim \mathbb{P}(T_i|Z_i,X_i)$, where T_i^j and $\hat{T}_i^j$ are the treatment and the predicted treatment for the j-th dimension of unit i. Afterwards, a balanced representation $f_\theta(X)$ is learned for continuous T via mutual information (MI) minimization constraints. Finally, CB-IV regresses Y with $\hat{T} \sim \mathbb{P}(T|Z,X)$. **CBRL-CIV [20]** divides the observed covariates X into CIVs and confounders, respectively denoted as S and C. The first step is to learn the representation $Z = \psi_\theta(C)$ so that $S \perp\!\!\!\perp Z \mid C$, eliminating the confounding bias between S and T. The second step is to regress T with $\{S,C\}$ without the influence of U. The last step is to regress Y with balanced representation Z and estimated treatment $\hat{T}$. **GCB-IV [146]** is an extension of CB-IV, which is specifically designed to handle cases without any predefined IVs. It uses VAE to recover latent variables $L = \{U_{IV}, U_X\}$ from observational data, and recover conditional probability distribution $\mathbb{P}(L|X,T)$. This module can help treatment regression by $\mathbb{P}(T|L)$, naturally plugged into the CV-IV framework.

5.2.3 Difference-in-Difference Methods. How to extend the DID methods of binary treatment into the setting of continuous treatments have been discussed [18] as well. Denoting the dosage as D, the *parallel trends* for continuous treatments are rewritten as $\mathbb{E}[Y_{g=1}(0) - Y_{g=0}(0)|D=d] = \mathbb{E}[Y_{g=1}(0) - Y_{g=0}(0)|D=0]$ for all $d \in \mathcal{D}$. The subscript g is consistent with the time indicator in Section 3.2.3. However, it does not involve potential outcome paths between treatment dosages, introducing selection bias if comparing outcome changes between different dosage groups. Therefore, ADRF is proved theoretically unidentifiable. A *strong parallel trends assumption* is proposed that $\mathbb{E}[Y_{g=1}(d) - Y_{g=0}(0)] = \mathbb{E}[Y_{g=1}(d) - Y_{g=0}(0)|D=d]$ for all $d \in \mathcal{D}$, excluding the systemic differences arising from the selection of different treatment dosages. After derivation, $ADRF(d) = \frac{\partial\mathbb{E}[\Delta Y|D=d]}{\partial d}$. This method adapts to different model structures in a data-driven manner instead of relying on assumption of function forms. The nonparametric estimator converges at the minimax rate and provides data-driven uniform confidence intervals with correct asymptotic coverage and adaptivity.

5.3 Conclusion and Discussion

Key challenge in the setting of continuous treatments is to solve the countless possible values of T. GPS is further developed here, which is modeled using Gaussian distribution to ensure its continuity. The problem of model misspecification in this scenario will be even worse. The goal of representation learning methods is to study a generally balanced embedding of X for all $T \in \mathcal{T}$. An intuitive solution is discretization of T into discrete intervals, as applied in DRNet and SCIGAN, but it may lead to jump discontinuity at the interval boundaries. Another solution is to learn common embeddings by a general module, such as varying coefficient model and transformer. Varying coefficient model [22] proposes a component-wise smoothing spline method for non-parametrically estimating the coefficient functions in varying coefficient models. It addresses the challenge of continuous treatments by allowing the regression coefficients to smoothly vary as a function of time. Through its use of smoothing splines, it inherently provides a continuous estimation of the effects, thus avoiding the discontinuities that arise from the discretization of a continuous variable. Transformers [137], on the other hand, are deep learning architectures based on self-attention mechanisms, originally used in natural language processing. Their powerful sequence modeling capabilities allow them to process complex continuous input data,

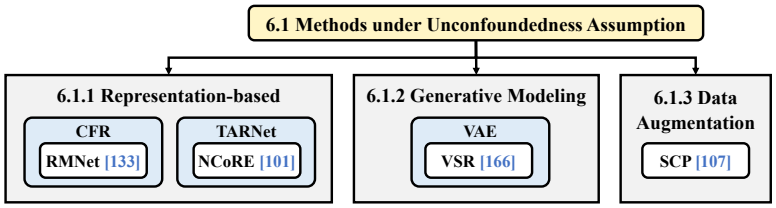


Fig. 7. Categorization of bundle treatment methods without unobserved confounders.

and through attention mechanisms, they can capture dynamic and nonlinear relationships between covariates and outcomes across the entire continuous treatment spectrum, thereby learning common, balanced representations. The methods introduced in this section to tackle unobserved confounders can be commonly used in both binary and continuous settings. Advantages and limitations summarized in Section 3.3 are still applicable for methods in the current setting.

6 Bundle Treatment

Different from the setting of multi-valued treatment $T \in \mathcal{T}^{mul} \subset \mathbb{R}$, a unit can simultaneously adopt several treatments $T \in \mathcal{T}^{bun} \subset \{0,1\}^m$ in the case of bundle treatment. Studies on this setting can also be divided into two categories, according to the existence of unobserved confounders.

6.1 Under Unconfoundedness

Methods related to bundle treatment and in accordance with the *unconfoundedness assumption* are included in Fig. 7. The majority of them belong to representation-based solutions or generative modeling methods. Single-cause perturbation method [107] explores a new way for counterfactual estimation of bundle treatment via data augmentation. Specifically, it obtains new counterfactual outcomes based on the existing treatment vector (e.g., $T = \{0,1,1,1,0\}$), by successively flipping a single treatment dimension (e.g., $T = \{0,1,1,1,1\}$). This simplifies the problem of addressing multi-dimensional treatments into that of addressing single treatment.

6.1.1 Representation-based Methods. CFR or TARNet are also developed into the setting of bundle treatment. One challenge is the more complex confounding bias, since the possible treatment space expands in an exponential manner, and a shared hypothesis network is thus needed for all the treatment groups with the purpose of sample efficiency. Moreover, the additional influence caused by interactions among treatments taken in the same time should also be considered.

Regret Minimization Network (RMNet) [133] is proposed to address the problem of sample efficiency, together with the gap between the regression accuracy for the whole treatment space and the decision-making performance with respect to an exact treatment. Decision-focused risk is proposed to mitigate the aforementioned gap, which is essentially a classification task to predict whether a treatment is assuredly better in term of the decision-making performance. Formally,

$$\widetilde{\mathrm{ER}}_g^u(f) = \mathop{\mathbb{E}}_X \left[-\frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \{S \log V + (1-S)\log(1-V)\} \right], \tag{37}$$

where $S := I(Y - g(X) \geq 0)$, $V := \sigma(f(X,T) - g(X))$, and $\sigma(\cdot)$ is the sigmoid function. The superscript u indicates it is a uniform metric different from the normal definition. Considering that the regression error (MSE) still plays an important role on decision-making, loss function with respect to the hypothesis network $h(\Phi)$ is formally defined as $\mathcal{L}^u(f;g) = \sqrt{\widetilde{\mathrm{ER}}_g^u(f) \cdot \mathrm{MSE}^u(f)}$. As for the challenge of sample efficiency in the representation network Φ, embeddings of x and t are both learned for the following inference. There are two alternative plans,

and the first is to construct a single network Φ to learn the joint representation and utilize IPM as a restriction. The second method is to construct two separate networks Φ_x and Φ_t and regularize them to be independent from each other by minimizing the Hilbert-Schmidt Independence Criterion (HSIC) [38].

Neural Counterfactual Relation Estimation (NCoRE) [101] makes attempts to model cross-treatment interactions by simulating their combined effects via the additive mechanism of neural network layers. It constructs arms for all single treatments, like the the multiple heads of $h(\Phi)$ in TARNet. The difference is that there is an merge layer connecting all the treatment arms in the end. When training the model, all the samples with the information of X will go through the base layers, while only the arms corresponding to those treatments involved in the bundle treatment will be updated. In the stage of prediction, the merge layer receives the outputs from related treatment arms as inputs and finally calculates the potential outcome.

6.1.2 Generative Modeling Methods. Bundle treatment can be regarded as a high-dimensional vector, and **Variational Sample Re-weighting (VSR) [166]** points out that it is feasible to learn a latent representation Z for T using VAE, and then decorrelate the low-dimensional Z with confounders X. Specifically, the objective function of VAE is to maximize:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z\sim q_\phi(Z|T_i)}\left[\log_{\varphi(T_i|Z)}+\log\mathbb{P}(Z)-\log q_\phi(Z|T_i)\right] \tag{38}$$

To remove the confounding bias, variational sample weight $w^d=\{w_i^d\}_{i=1}^n$ is also proposed as:

$$w_i^d=W_T(X_i,T_i)=\frac{\mathbb{P}(T_i)}{\mathbb{P}(T_i|X_i)}=\frac{1}{\mathbb{E}_{z\sim q_\phi(Z|T_i)}\left[\frac{1}{W_Z(X_i,Z)}\right]}, \tag{39}$$

where $W_Z(X,Z)$ is the density ratio estimation with which T can be decorrelated with X. Specifically, the data points from observational dataset are regarded as positive samples ($L=1$) while those from decorrelated target dataset are negative samples ($L=0$). In this way, we define:

$$W_Z(X,Z)=\frac{\mathbb{P}(X,Z|L=0)}{\mathbb{P}(X,Z|L=1)}=\frac{\mathbb{P}(L=1)}{\mathbb{P}(L=0)}\cdot\frac{\mathbb{P}(L=0|X,Z)}{\mathbb{P}(L=1|X,Z)}=\frac{\mathbb{P}(L=0|X,Z)}{\mathbb{P}(L=1|X,Z)}. \tag{40}$$

There is a classifier $p_{\theta_d}(L|X,Z)$ to give the values of $\mathbb{P}(L|X,Z)$ with the limitation that $\frac{\mathbb{P}(L=1)}{\mathbb{P}(L=0)}=1$ for all the data points. Finally, a network $f_{\theta_p}(X_i,T_i)$ is learned to predict the potential outcome.

6.1.3 Data Augmentation. **Single-cause Perturbation (SCP) [107]** provides a new insight that the counterfactual prediction for bundle treatment can be achieved via data augmentation. In other words, SCP perturbs a single treatment (M in total) to be its opposite value, thus generating additional data by predicting the potential outcomes. In this way, the treatment assignment becomes more balanced than the original observational data, which mitigates the confounding bias in an entirely new way. Under the *sequential ignorability assumption* [110], conditional expectation of the potential outcome for a single treatment has equivalence to that of a bundle treatment:

$$\mathbb{E}\left[Y(t_m,t_{-m})|X\right]=\mathbb{E}\left[Y(t_m)|X,T_{-m}(t_m)=t_{-m}\right], \tag{41}$$

where t_m is the m-th treatment in the bundle, while t_{-m} denotes the remaining ones. This assumption plays a key role for the validity of SCP, simplifying the problem of effect estimation for bundle treatment into that for a single treatment. Important modules include single-cause model training, data augmentation, and covariate adjustment. The goal of the first module is to well estimate the holistic potential outcome $\mathbb{E}[Y|X'_m,T^{\downarrow}_{-m},T_m]$, where two estimators are need for a single treatment t_m and its causal descendants $T^{\downarrow}_{-m}(t_m)$, respectively. Disentangled Representations for Counterfactual Regression algorithm (DR-CFR) [46] is applied here as the estimators. It is able to sample perturbed data points for the data augmentation once the single-cause model is fitted. The perturbed dataset is $\mathcal{D}_m=\{x_i,\tilde{y}_i^m,\tilde{t}_i^m\}_{i=1}^n$. The newly generated data for all the single treatments are merged

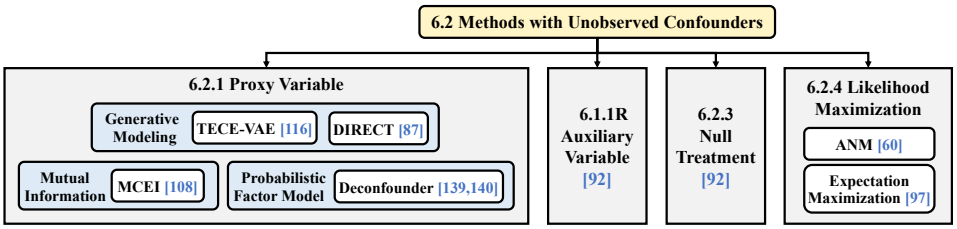


Fig. 8. Categorization of bundle treatment methods with unobserved confounders.

together as the augmented dataset. As for the covariate adjustment, a standard feed-forward neural network $f_\theta : \mathbb{R}^D \times \Omega \to \mathbb{R}$ is trained on the augmented dataset to learn $\mathbb{E}[Y(t)|X = x] = \mathbb{E}[Y|X = x, T = t]$.

6.2 With Unobserved Confounders

The problems of unobserved confounders have also been studied under the circumstance of bundle treatment. The relevant methods are reviewed in Fig. 8.

6.2.1 Proxy Variable. **TECE-VAE [116]** expands CEVAE to the bundle treatment setting, and introduces task embedding to model the interdependence among multiple treatments. It allows a flexible representation of a task by multiplying a vector of zeros and ones, meaning which treatments are applied, and a weight matrix W is learned. A network g_1 is trained to learn $\mathbb{P}(T|X)$, from which the treatment vector $\tilde{T}$ is sampled. Multiplying $\tilde{T}$ by the embedding matrix W gives the new representation $\tau = W\tilde{T}$. Another network g_2 is trained to learn $\mathbb{P}(Y|T, X)$ given τ, where the potential outcome $\tilde{Y}$ is sampled. Networks g_3 and g_4 give the mean and variance that determine the conditional distribution of latent confounders V, i.e., $\mathbb{P}(V|T, X, Y)$. The purpose of the decoder is to reconstruct $\{X, T, Y\}$, with the input V sampled from $\mathbb{P}(V|T, X, Y)$. Networks $f_1 \sim f_4$ are established for various types of X (binary, categorical, or continuous). f_5 is similar to g_1, which outputs $\mathbb{P}(T|V)$ for sampling the treatment vector $\tilde{T}$. The new representation τ is obtained in the same way as described in the encoder. Ultimately, f_6 aims to learn $\mathbb{P}(Y|T, V)$, determining Y given V and τ.

Disentangled Multiple Treatment Effect Estimation (DIRECT) [87] aims to learn the representation of confounder proxy V from the treatment assignments by VAE, and further explores the interdependence of multiple treatments. There are two main blocks including an inference network and a generation network. The objective of the inference network is to learn the disentangled representation of V. To be specific, the embeddings of every single treatment T_j are learned according to the treatment assignment A, followed by a clustering module $f_c(\cdot)$ to approximately simulate the distribution of each class, i.e., $\mathbb{P}(C_j|T_j) = Mult\left(f_c(t_j)\right)$. Notation C_j here represents the cluster assignment of T_j, and $Mult(\cdot)$ is Multinomial distribution. Such idea is similar to VaDE [61]. Afterwards, disentangled confounder representation $V^{(k)}$ is learned for each class, which is implemented in a manner similar to β-VAE [48]. The holistic confounder representation of V is obtained by concatenating all of them. In the generation network, the main task is to reconstruct the treatment assignment A when given the treatments T along with V. Moreover, the observational outcomes are also used as supervision for better capturing the latent confounders, and the prediction loss is defined as $\mathcal{L}_y = -\sum_{i=1}^{n} \log \mathbb{P}(\hat{Y}_i = Y_i|V_i, A_i, T)$. The overall loss function of DIRECT is:

$$\begin{aligned}\mathcal{L} = &- \mathbb{E}[\log \mathbb{P}(A|V, T, C)] + \mathbb{E}_{q(T|A)} KL\left(q(C|T) \| \mathbb{P}(C)\right) + \mathbb{E}_{q(C|T)} KL\left(q(T|A) \| \mathbb{P}(T|C)\right) \\ &+ \lambda \cdot \mathcal{L}_y + \beta \cdot \sum_{k=1}^{K} \mathbb{E}_{q(T|A)q(C|T)} KL\left(q\left(V^{(k)}|A\right) \| \mathbb{P}\left(V^{(k)}\right)\right),\end{aligned} \tag{42}$$

where hyper-parameters β and λ are used to control the effect of different parts.

Multiple Causal Estimation via Information (MCEI) [108] aims to recover proxy variable V by constraints of mutual information. Two additional assumptions are proposed in the case of multi-valued treatments. The first is *shared confounding assumption* that the confounders are shared across all treatments, and thus each treatment could reflect some information of the shared confounders. The second is *independence given unobserved confounders*, meaning that treatments are independent given latent confounder V so as to avoid the dependencies between T_i and T_{-i}, where T_{-i} denotes the set of treatments excluding the i-th treatment. A confounder estimator for V is established to reconstruct a treatment T_i when given the remaining ones T_{-i}. Additional Mutual Information (AMI) is utilized to measure how much additional information could T_i provides for the confounders. After confounder V is recovered, the potential outcome can be modeled with the residuals and confounder by maximizing the following objective:

$$\max_{\eta} \mathbb{E}_{\mathbb{P}(Y,T)p_\theta(U|T)\mathbb{P}(\xi_i|U,T)}[\log\ p_\eta(Y|U,\xi)], \tag{43}$$

where θ is the parameter associated with the estimation of latent confounder U, ξ_i denotes the independent component of the i-th treatment, $p_\eta(Y|U,\xi)$ is the outcome model with parameter η.

Another way to capture latent confounders of multiple treatments is using probabilistic factor model. **Deconfounder (2018) [139]** points out that a variable making all the treatments conditionally independent from each other could be found, once a factor model well representing the treatment distribution is figured out. Details of implementations can be concluded as:

(1) Finding out a suitable model for latent variable according to the treatment assignment, namely fitting a probabilistic factor model to capture the joint distribution among them:

$$U_i \sim p(\cdot|\alpha) \quad i = 1,\ldots,n \text{ , and } T_{ij}|U_i \sim p(\cdot|v_i,\theta_j) \quad j = 1,\ldots,m \tag{44}$$

where α refers to the parameters for distribution of U_i, and θ_j denotes those for the per-cause distribution of T_{ij}. Note that i is the index for each sample while j is that of each cause.

(2) Inferring the latent variable for each sample through $\hat{U}_i = \mathbb{E}_M[U_i|T_i = t_i]$.

(3) Estimating the casual effect by utilizing $\hat{U}_i$ as a substitute of the confounders:

$$\mathbb{E}[Y_i(t)] = \mathbb{E}\left[\mathbb{E}\left[Y_i(t)|\hat{U}_i, T_i = t\right]\right]. \tag{45}$$

Deconfounder (2021) [140] is an improved version of Deconfounder (2018). There are some key findings that a subset C of treatments could be regarded as proxies of the unobserved confounders, helping the identification for the remaining treatments. The distribution of C can be determined as well. Implementations are similar to those of Deconfounder (2018), which are described as follows.

(1) Constructing latent variable $\hat{U}$ that makes all the treatment conditionally independent by $\hat{\mathbb{P}}(T_1,\ldots,T_m,\hat{U}) = \hat{\mathbb{P}}(\hat{U})\prod_{j=1}^{m}\hat{\mathbb{P}}(T_j|\hat{U})$, where $\hat{\mathbb{P}}(\cdot)$ is consistent with observational data, i.e., $\mathbb{P}(T_1,\ldots,T_m) = \int \hat{\mathbb{P}}(T_1,\ldots,T_m,\hat{U})d\hat{U}$.

(2) Fitting the outcome model $\mathbb{P}(Y,T_1,\ldots,T_m)$ by $\int \hat{\mathbb{P}}(Y|T_1,\ldots,T_m,\hat{U})\hat{\mathbb{P}}(T_1,\ldots,T_m,\hat{U})\ d\hat{U}$.

(3) Estimating $\hat{\mathbb{P}}(Y|\mathrm{do}(t_C)) \triangleq \int \hat{\mathbb{P}}(T_1,\ldots,T_m,\hat{U}) \times \hat{\mathbb{P}}(T_{\{1,\ldots,m\}\setminus C},\hat{U})\ d\hat{U}\ dT_{\{1,\ldots,m\}\setminus C}$.

6.2.2 Auxiliary Variable. Similar to proxy and IV, auxiliary variable [92] has no causal relationship with Y according to the following assumptions. The first is *exclusion restriction*, i.e., $Z \perp\!\!\!\perp Y|(X,U)$. Two additional assumptions are proposed to limit the joint distribution between T and U:

(1) *Equivalence.* For any α, any $\tilde{f}(x,u|z)$ that solves $f(x|z;\alpha) = \int_u \tilde{f}(x,u|z)du$ can be written as $\tilde{f}(x,u|z) = f\{X = x, v(U) = u|z;\alpha\}$ for an invertible but not necessarily known function v.

(2) *Completeness.* For any α, $f(u|x,z;\alpha)$ is complete in z, i.e., for any fixed x and square-integrable function g, $E\{g(U)|X = x, Z;\alpha\} = 0$ almost surely if and only if $g(U) = 0$ almost surely.

Equivalence is a high-level assumption stating that the treatment-confounder distribution lies in a model that is identified upon a one-to-one transformation of U. Because the *unconfoundedness assumption* holds conditional

on any one-to-one transformation of U, this allows us to use an arbitrary admissible treatment-confounder distribution to identify the treatment effects. *Completeness* is a fundamental concept in statistics, meaning that conditional on X, any variability in U is captured by variability in Z, analogous to the relevance condition in the instrumental variable identification. When both U and Z have k levels, completeness means that the matrix $[f(u_i|x, z_j)]_{k\times k}$ consisting of the conditional probabilities is invertible. Under these assumptions, identification with auxiliary variable is proved feasible. Algorithm is described as follows.

(1) Obtaining an arbitrary admissible joint distribution $\tilde{f}(x, u, z)$.
(2) Solving $f(y|x, z) = \int_u \tilde{f}(y|u, x)\tilde{f}(u|x, z)du$ with $\tilde{f}(u|x)$ from step (1) and estimated $f(y|x)$.
(3) Plugging the estimation of $\tilde{f}(y|u, x)$ from Step (2) and the estimation of $\tilde{f}(u)$ derived from $\tilde{f}(u, x, z)$ into $f\{Y(x) = y\} = \int_u \tilde{f}(y|u, x)\tilde{f}(u)du$, thus estimating $f\{Y(x)\}$.

6.2.3 Null Treatment Method. This method also depends on the *equivalence assumption* and *completeness assumption* mentioned above. Another key assumption called *Null treatment* [92] is proposed as well. The cardinality of the intersection $C \cap \mathcal{A}$ does not exceed $(|C| - q)/2$, where $|C|$ is the cardinality of C and must be larger than the dimension of U. Implementations conclude:

(1) Obtaining an arbitrary admissible joint distribution $\tilde{f}(x, u)$.
(2) Using the estimate $\tilde{f}(u|x)$ from Step (1), along with an estimation of $f(y|x)$, to solve $f(y|x) = \int_u \tilde{f}(y|u, x)\tilde{f}(u|x)du$ for $\tilde{f}(y|u, x)$.
(3) Plugging the estimate of $\tilde{f}(u)$ from Step (1) and $\tilde{f}(y|u, x)$ from Step (2) into $f\{Y(x) = y\} = \int_u \tilde{f}(y|u, x)\tilde{f}(u)du$ to estimate $f\{Y(x)\}$.

6.2.4 Likelihood Maximization. Researchers also attempt to estimate the effect of bundle treatment in the presence of hidden confounders under the framework of Structural Causal Model (SCM). They have proved that it is impossible to identify the joint effects of multiple treatments that are simultaneously taken if there is no restriction on the structure function [97]. However, such influence could be estimated by introducing reasonable assumptions, such as *additive noise model (ANM)*. For example, an ANM-based method [60] is proposed, where all data from different regimes are pooled together to jointly maximize the combined likelihood. A complementary question is also studied that how to estimate the causal effect of a single treatment while multiple treatments are adopted at the same time [60]. Formally, given the samples that can deduce $\mathbb{E}[Y|T_i = t_i, T_j = t_j, X = x]$ and $\mathbb{E}[Y|do(T_i = t_i, T_j = t_j), X = x]$, the purpose is to find how to learn the conditional average treatment effect $\mathbb{E}[Y|do(T_i = t_i), T_j = t_j, X = x]$ or $\mathbb{E}[Y|T_i = t_i, do(T_j = t_j), X = x]$. Researchers prove that this is not generally possible as well, unless there are non-linear continuous structural causal models with additive, multivariate Gaussian noise. They extend the Expectation Maximization style iterative algorithm [97] to disentangle the effects of each single treatment. Suppose the intervened treatments are $T_{int} \subseteq T$ and $T_{obs} \equiv T - T_{int}$, and then a causal query with T_{int} could be decomposed as $\mathbb{E}[Y|C; do(X_{int}); X_{obs}] = f_Y(C; X) + \mathbb{E}[U_Y|X_{obs}]$.

6.3 Conclusion and Discussion

Under *Unconfoundedness Assumption*, the key challenges in this setting include: (1) complex confounding bias for exponential-level treatment assignments, (2) the need for a general hypothesis model for all treatment groups, and (3) additional influence caused by the interactions of multiple treatments that are simultaneously taken.

7 Method Extension and Assumption Verification

7.1 Method Extension

Propensity Score. It is defined as the conditional probability of individual receiving a treatment when given the covariate. In the setting of binary treatment, PS is $e(X) = \mathbb{P}(T = 1|X)$ and usually modeled by logistic regression.

When treatment becomes multi-valued, the generalized propensity score (GPS) is extended from PS and denoted as $e(T, X) = f_{T|X}(T, X)$. It is similar to a multi-label classification task, and usually modeled by multinomial logistic regression. As for the continuous treatment, the definition of GPS remains unchanged but Gaussian distribution is used instead for its continuity property. Bundle treatment can be regarded as a multi-dimensional vector, and the distribution $f_{T|X}(T, X)$ can be learned by neural networks in this setting. Once these propensity scores are quantified, the matching, stratification, and re-weighting methods can be naturally extended to the complex treatment environments. However, relying solely on GPS may not be sufficient to capture the complex causal relationships when treatments are continuous or bundle. It is more like serving as auxiliary information in counterfactual prediction or covariate balancing.

Covariate Balancing and Representation Learning. In the binary treatment setting, covariate balance between different treatment groups can be directly achieved through re-weighting, with various constraints such as first-order moment alignment. However, such sample weight works by $\tilde{X}_i = w_i X_i$, treating all the confounders (each dimension in X) equally in a linear manner. Therefore, representation learning network Φ is applied for balanced covariate embedding $\Phi(X)$. Besides remaining the helpful information for predicting Y, the most important property of $\Phi(X)$ is balance between different groups. Many works, such as BNN and CFR, use IPM or MMD to constrain the alignment between $\mathbb{P}(X|T = 0)$ and $\mathbb{P}(X|T = 1)$. When T is optional from m discrete values (multi-valued) or possible combinations (bundle), such structure can be intuitively extended as one common representation network Φ with C_m^2 IPM or MMD constraints and m prediction heads to respectively learn $\mathbb{P}(Y|X, T = t)$ for $t \in \{1, \ldots, m\}$, like methods described in Section 4.1.3 and Section 6.1.1. Even for the continuous treatment, this structure is applicable as long as the whole space $\mathcal{T}^{con}$ is discretized into m intervals, taking DRNet for example. However, considering the issue of continuity, representation learning for continuous treatment may involve using coefficient varying models or substitute multiple distribution distance constraints with mutual information.

GAN. The framework using GAN is a general one for different treatments, where the generator G is designed for targets of interests and the discriminator D is for alignment. For example, GANITE predicts the counterfactual outcomes of different treatments by G and ensures its effectiveness by letting D judge which treatment corresponds to the factual data. It essentially enforces the alignment of counterfactual data to real data in G via being attacked by D. Since D is designed as a classification task, the framework of GANITE is naturally applicable for treatments with limited action space, i.e binary, multi-valued and bundle treatments. As for SCIGAN, applying a similar hierarchical structure in DRNet, it succeeds to utilize the GAN framework for continuous treatment through discretization. The continuity property of response function has not been ensured. Moreover, these methods have no theoretical proof for the identifiability of ITE or IDRF.

Latent variable recovery with VAE. Under the structure with encoder and decoder, VAE is able to recovery the latent variables Z as proxies by capturing and deconstructing the relationships between Z and other observed variables. Whatever the treatment type, VAE is applicable once there exist helpful latent variables. VAE here is more like a general technology rather than a specific framework. The real challenge lies in confirming the existence of latent proxies and how to utilize them to facilitate treatment effect estimation.

7.2 Assumption Verification and Relaxation

As stated in Section 2, three assumptions are typically required [47, 56] in causal inference. We also provide an illustration in Fig. **??**.

- *SUTVA*. It has two parts. First, *no interference* means one subject's treatment does not affect another's outcome. We can test this using methods such as A/B testing [41, 106, 123], which involve controlling neighbors' treatment status to check for spillover effects. Second, *treatment consistency* assumes each treatment has a single, unambiguous interpretation (e.g., a specific drug or dosage) [56]. This is usually

met in most practical settings. For instance, in a binary choice scenario, if the treatment is "taking drug A" versus "not taking drug A," there is usually no ambiguity about what "taking drug A" means. Similarly, for a continuous intervention like a drug's dosage, a dosage of "10mg" is clearly distinct from "20mg".

- *Overlap*. This means there is a non-zero probability for any subject to receive any of the treatments being studied, given their characteristics. We can assess overlap by examining propensity scores or using tools like kernel density estimation to visualize the distribution of different treatment groups [26, 30].
- *Unconfoundedness*. This is the crucial assumption that all common causes of treatment and outcome (confounders) have been measured and accounted for. Since this can be hard to guarantee, sensitivity analysis is widely used to explore how robust results are to potential hidden confounders [47, 162]. Alternatively, methods using instrumental variables (IVs), negative controls (NCs), or exploiting multiple environments can help identify or address the impact of unmeasured confounders [64, 65].

To relax the *unconfoundedness assumption*, specific causal inference methods often rely on data generation assumptions or proxy variables:

- *Data generation assumptions*. These involve specific characteristics of how the data was produced, such as *additive noise* [11, 45, 126] (where noise is simply added to the causal effect) or *parallel trends* [6, 8, 18] (common in difference-in-differences designs, assuming outcomes would have followed similar trends without intervention). These assumptions are difficult to validate in real-world scenarios, researchers have to rely on prior knowledge about the data generation process itself to study cases.
- *Proxy variables*. These leverage additional variables in the observational data to help relax the unconfoundedness assumption. This includes assumptions related to instrumental variables and proxy variables, which must satisfy specific independence conditions. For IVs, the Sargan–Hansen test can help validate their relevance and exogeneity [119]. For negative controls, subject matter knowledge is often crucial for their appropriate selection and validation [126].

Although there are some studies on relaxing or violating *SUTVA* [25, 52, 149, 161] and *overlap* [57, 59, 90, 150], they mainly focus on the scenario of binary treatment. Their extension to the setting of complex treatments still remains an open problem.

8 Datasets and Codes

8.1 Available Datasets

We summarized the datasets applicable for evaluating methods of complex treatments in Table **??**. For multi-valued treatments, Twins and News are available.

- **Twins.** This dataset is collected from all births in the USA between 1989-1991, and only the twins weighing less than 2kg are recorded without missing features [3]. The outcome refers to the mortality after one year. It is originally utilized by models focused on binary treatment. After preprocessing, there are 11, 400 pairs of twins, along with 30 covariates related to the parents, pregnancy and birth. It can be extended to multi-valued setting, where 4 treatments are considered: $T = 0$ refers to samples with lower weight and female sex, $T = 1$ means those with lower weight and male sex, $T = 2$ is those with higher weight and female sex, and $T = 3$ means those with higher weight and male sex.
- **News.** It simulates the opinions of a media consumer exposed to multiple news items, which is generated from the NY Times corpus. The purpose is to infer the individual treatment effects of obtaining more content from some specific devices on the reader's opinions. In particular, each sample x_i refers to news items represented by word counts, and outcome $y_i \in \mathbb{R}$ represents the reader's opinions of the news. Multiple treatments can be constructed as various devices used to view the news, such as desktop, smartphone, and newspaper.

There are 4 more datasets used in the case of continuous treatment.

- **Cancer Genomic Atlas (TCGA) [142].** It contains 9, 659 observations with 20, 531 features from various cancers. Three clinical treatments are taken into account including medication, chemotherapy, and surgery. The outcome studied here is the risk of cancer recurrence.
- **Medical Information Mart for Intensive Care (MIMIC) III [63]**. It is a large database recording 8, 040 patients who are admitted to ICUs at a large tertiary care hospital. Beside the 49 features of a patient, it also includes a wide range of clinical data, including demographic information, vital signs, laboratory test results, medications, and clinical notes.
- **Infant Health and Development Program (IHDP) [15]**. It is a longitudinal study that was conducted in the United States from 1985 to 1993. It contains data from 747 infants with 25 covariates. The infants were randomly assigned to either a treated group with high-quality educational and developmental services, or a control group with standard care.
- **Medicare [151]**. This is a collection of data on socioeconomic status for 2, 132 US counties, together with average annual cardiovascular mortality rate (CMR) and total PM 2.5 concentration. This study spans over 21 years (1990-2010) and covers more than 68.5 million samples. Several works [109, 152] focus on this dataset to estimate the long-term causal effect of PM 2.5 on all-cause mortality under 18 covariates.

We summarize the available real-world or semi-synthetic datasets for bundle treatment.

- **Smoking [54].** The 1987 National Medical Expenditures Survey (NMES) collected data on smoking habits and medical expenses in a representative sample of the U.S. population. The dataset contains 9, 708 people and 8 variables about each.
- **TMDB 5000 Movie Dataset.** It is a collection in Kaggle that contains 901 actors (who appeared in at least 5 movies) and the revenue for the 2, 828 movies they appeared in. The movies span 18 genres and 58 languages. This dataset can be used to study how much does an actor boost (or hurt) a movie's revenue.
- **Amazon-3C** and **Amazon-6C.** They are two semi-synthetic datasets from the Amazon review dataset in DIRECT [87]. In each dataset, 3/6 categories of items are selected. Afterwards, the top 1, 000 products with most reviews are collected as instances in each category. The goal is to investigate the effect of the keywords in reviews on the future sales of each product. Specifically, treatments here refer to 3 key words derived from the reviews, and the outcome is future amount of sales for each product. Confounders are the latent attributes of the products.
- **CRISPR Thee-way Knockout (CRISPR KO) [164].** It is a benchmark dataset collected in a systematic multi-gene knockout screen. It is particularly challenging because of the complex underlying biological process, which leads to large number of potential treatment combinations, high-dimensional covariates, and sparsity of labeled data.

8.2 Open-source

The available codes for causal inference with complex treatments are summarized in Table **??**, including multi-valued, bundle, and continuous settings. Considering that the IV methods can be naturally developed to solve the causal estimation of continuous treatment, readers can also refer to the toolkit[2] of IV methods that is reviewed in the survey of IV [147].

9 Future Directions

Despite substantial progress, the study of causal inference with complex treatments still confronts several challenges that require further exploration.

[2]https://github.com/causal-machine-learning-lab/mliv

- **Confounding heterogeneity**. Existing bundle treatment methods typically assume uniform confounding across all treatments—an assumption often violated when individual treatments are subject to distinct confounding mechanisms, leading to biased effect estimates. Flexible deconfounding techniques that accommodate treatment-specific confounding structures remain an important open direction.
- **Treatment interactions**. Quantifying interaction effects—synergistic, antagonistic, or merely additive—among concurrent treatments is critical for optimal intervention design in numerous practical domains. However, developing interpretable and scalable estimation methodologies remains challenging, especially under complex confounding and high-dimensional settings.
- **Data sparsity and vast treatment spaces**. Available data typically cover only a tiny fraction of possible treatment combinations or dosage levels. Extrapolation across enormous treatment spaces can be highly vulnerable to selection bias [74–77] and *overlap* violations [19, 135]. Effective exploration and optimization of the full treatment landscape demand innovative strategies such as adaptive trial designs, reinforcement learning, and transfer learning.

Moving forward, several emerging machine learning paradigms present a promising path for tackling these core challenges and advancing causal inference with complex treatments.

- **Advanced generative models for counterfactual inference**. Beyond the GAN- and VAE-based approaches discussed in Section 7.1, diffusion models and other modern generative architectures are increasingly employed for counterfactual outcome synthesis [58, 70, 118]. Their training stability and higher-fidelity generation make them especially attractive for continuous treatments and high-dimensional (bundle) settings.
- **Integration of large language models (LLMs)**. LLMs are increasingly used as "world simulators" to generate counterfactual scenarios, extract causal structures from text, or even assist in experimental design [28, 67, 79, 86, 132, 148]. This paradigm holds particular promise for settings rich in natural-language data but poor in conventional covariates.
- **Causal representation learning for complex treatments**. For complex interventions (e.g., bundle treatments), identifying causal representations, which remain invariant under interventions on subsets of variables, is an essential prerequisite. Recent advances in causal representation learning [1, 17, 80, 100, 120] provide the foundational progress needed to achieve reliable and generalizable causal estimation.

10 Conclusion

The primary goal of causal inference is to determine the difference in outcomes for a population if they had received a particular treatment compared to if they had not. It has extensive applications in fields such as epidemiology, finance, and education. However, in many of these real-world scenarios, treatments are multi-valued, continuous, bundle, or even more complex. In this survey, we provide a comprehensive review of methods tackling such complex treatments. We first introduced the basic notations, concepts, and assumptions. Then we introduced the relevant methods according to the type of treatment and the presence of unobserved confounders. We also discussed the intrinsic connections of these methods and the required assumptions. Available datasets and open-source code are also reviewed. Finally, some potential directions are discussed for future explorations.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (2024YFE0203700), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037), and the National Natural Science Foundation of China (62376243). All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. 2023. Interventional causal representation learning. In *International conference on machine learning*. PMLR, 372–407.

[2] Sina Akbari and Negar Kiyavash. 2024. Non-linear Triple Changes Estimator for Targeted Policies. *arXiv preprint arXiv:2402.12583* (2024).

[3] 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.

[4] Daniel Arbour, David Dimmery, and Akshay Sondhi. 2021. Permutation Weighting. In *International Conference on Machine Learning*. PMLR, 331–341.

[5] Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. Synthetic Difference-in-Differences. *American Economic Review* 111, 12 (December 2021), 4088–4118.

[6] Orley Ashenfelter. 1978. Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics* 60, 1 (1978), 47–57.

[7] Susan Athey, Guido Imbens, and Stefan Wager. 2016. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B* 80 (2016).

[8] Susan Athey and Guido W Imbens. 2006. Identification and Inference in Nonlinear Difference-In-Differences Models. *Econometrica* 74, 2 (2006), 431–497.

[9] TA Bancroft. 1944. On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics* 15, 2 (1944), 190–204.

[10] Andrew Bennett and Nathan Kallus. 2020. The Variational Method of Moments. *CoRR* abs/2012.09422 (2020).

[11] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep Generalized Method of Moments for Instrumental Variable Analysis. In *Advances in Neural Information Processing Systems*. 3559–3569.

[12] Ioana Bica, James Jordon, and Mihaela van der Schaar. 2020. Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*.

[13] Leo Breiman. 2017. *Classification and regression trees*. Routledge.

[14] Carlos Brito and Judea Pearl. 2002. Generalized Instrumental Variables. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*.

[15] Jeanne Brooks-Gunn, FR Liaw, and Pamela K Klebanov. 1992. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics* 120, 3 (1992), 350–359.

[16] Derek W Brown, Thomas J Greene, Michael D Swartz, Anna V Wilkinson, and Stacia M DeSantis. 2021. Propensity score stratification methods for continuous treatments. *Statistics in medicine* 40, 5 (2021), 1189–1203.

[17] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2023. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems* 36 (2023), 45419–45462.

[18] Brantly Callaway, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2024. Difference-in-Differences with a Continuous Treatment. arXiv:2107.02637

[19] Rui Chen, Guanhua Chen, and Menggang Yu. 2023. A generalizability score for aggregate causal effect. *Biostatistics* 24, 2 (2023), 309–326.

[20] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2024. Conditional Instrumental Variable Regression with Representation Learning for Causal Inference. In *International Conference on Learning Representations*.

[21] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

[22] Chin-Tsang Chiang, John A Rice, and Colin O Wu. 2001. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* 96, 454 (2001), 605–619.

[23] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2006. Bayesian Ensemble Learning. In *Advances in Neural Information Processing Systems*. 265–272.

[24] 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.

[25] Mayleen Cortez, Matthew Eichhorn, and Christina Yu. 2022. Staggered rollout designs enable causal inference under interference without network knowledge. *Advances in Neural Information Processing Systems* 35 (2022), 7437–7449.

[26] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. 2011. Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt*. Springer, 95–100.

[27] 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.

[28] Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. 2024. End-to-end causal effect estimation from unstructured natural language data. *Advances in Neural Information Processing Systems* 37 (2024), 77165–77199.

[29] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. 2020. Minimax Estimation of Conditional Moment Models. In *Advances in Neural Information Processing Systems*.

[30] Vassiliy A Epanechnikov. 1969. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14, 1 (1969), 153–158.

[31] M. Farrell, Tengyuan Liang, and S. Misra. 2020. Deep learning for individual heterogeneity: an automatic inference framework.

[32] Christian Fong, Chad Hazlett, and Kohsuke Imai. 2018. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics* 12 (03 2018), 156–177.

[33] Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. 2020. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 1 (2020), 199–214.

[34] Antonio F Galvao and Liang Wang. 2015. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *J. Amer. Statist. Assoc.* 110, 512 (2015), 1528–1542.

[35] Melissa M Garrido, Jessica Lum, and Steven D Pizer. 2021. Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings. *Statistics in medicine* 40, 5 (2021), 1204–1223.

[36] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[37] Sander Greenland. 1980. The effect of misclassification in the presence of covariates. *American journal of epidemiology* 112, 4 (1980), 564–569.

[38] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. 2007. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems*, Vol. 20.

[39] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2021. A Survey of Learning Causality with Data: Problems and Methods. *ACM Comput. Surv.* 53, 4 (2021), 75:1–75:37.

[40] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 20, 1 (2012), 25–46.

[41] Kevin Han, Shuangning Li, Jialiang Mao, and Han Wu. 2022. Detecting interference in a/b testing with increasing allocation. *arXiv preprint arXiv:2211.03262* (2022).

[42] Lars Peter Hansen. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 4 (1982), 1029–1054.

[43] Lars Peter Hansen and Kenneth J Singleton. 1982. Generalized instrumental variables estimation of nonlinear rational expectation models. *Econometrica* 50, 5 (1982), 1269–1286.

[44] Wolfgang Karl Härdle and Léopold Simar. 2019. *Applied multivariate statistical analysis*. Springer Nature.

[45] Jason S. Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A Flexible Approach for Counterfactual Prediction. In *International Conference on Machine Learning*, Vol. 70. 1414–1423.

[46] Negar Hassanpour and Russell Greiner. 2020. Learning Disentangled Representations for CounterFactual Regression. In *International Conference on Learning Representations*.

[47] M. A. Hernán and J. M. Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

[48] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

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

[50] Keisuke Hirano and Guido W Imbens. 2004. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164 (2004), 73–84.

[51] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

[52] Tadao Hoshino and Takahide Yanagi. 2024. Causal inference with noncompliance and unknown interference. *J. Amer. Statist. Assoc.* 119, 548 (2024), 2869–2880.

[53] Kosuke Imai and Marc Ratkovic. 2013. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76, 1 (07 2013), 243–263.

[54] Kosuke Imai and David A Van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* 99, 467 (2004), 854–866.

[55] Guido W. Imbens. 2000. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika* 87, 3 (2000), 706–710.

[56] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[57] Jaehyuk Jang, Suehyun Kim, and Kwonsang Lee. 2024. Improving Causal Estimation by Mixing Samples to Address Weak Overlap in Observational Studies. *arXiv preprint arXiv:2411.10801* (2024).

[58] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2022. Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision*. 858–876.
[59] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. 2020. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems* 33 (2020), 11637–11649.
[60] Olivier Jeunen, Ciarán M. Gilligan-Lee, Rishabh Mehrotra, and Mounia Lalmas. 2022. Disentangling Causal Effects from Sets of Interventions in the Presence of Unobserved Confounders. In *NeurIPS*.
[61] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2017. Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In *International Joint Conference on Artificial Intelligence*. 1965–1972.
[62] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. PMLR, 3020–3029.
[63] Alistair EW Johnson, Tom J Pollard, Lu Shen, Hung Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
[64] Rickard Karlsson and Jesse Krijthe. 2023. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems* 36 (2023), 44280–44309.
[65] Rickard KA Karlsson and Jesse H Krijthe. 2025. Falsification of Unconfoundedness by Testing Independence of Causal Mechanisms. *arXiv preprint arXiv:2502.06231* (2025).
[66] Prableen Kaur, Agoritsa Polyzou, and George Karypis. 2019. Causal Inference in Higher Education: Building Better Curriculums. In *Proceedings of the Sixth ACM Conference on Learning @ Scale, L@S 2019, Chicago, IL, USA, June 24-25, 2019*. 49:1–49:4.
[67] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research* (2023).
[68] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.
[69] Yaroslav Kivva, Saber Salehkaleybar, and Negar Kiyavash. 2023. A Cross-Moment Approach for Causal Effect Estimation. In *Conference on Neural Information Processing Systems*. 9944–9955.
[70] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. 2024. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. *arXiv preprint arXiv:2404.17735* (2024).
[71] 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.
[72] Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* 101, 2 (2014), 423–437.
[73] Greg Lewis and Vasilis Syrgkanis. 2018. Adversarial Generalized Method of Moments. *CoRR* abs/1803.07164 (2018).
[74] Baohong Li, Haoxuan Li, Anpeng Wu, Minqin Zhu, Shiyuan Peng, Qingyu Cao, and Kun Kuang. 2024. A Generative Approach for Treatment Effect Estimation under Collider Bias: From an Out-of-Distribution Perspective. In *International Conference on Machine Learning*. PMLR, 28132–28145.
[75] Baohong Li, Haoxuan Li, Ruoxuan Xiong, Anpeng Wu, Fei Wu, and Kun Kuang. 2024. Learning Shadow Variable Representation for Treatment Effect Estimation under Collider Bias. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 28146–28163.
[76] Baohong Li, Yingrong Wang, Anpeng Wu, Ming Ma, Ruoxuan Xiong, and Kun Kuang. 2025. Generalizing Causal Effects from Randomized Controlled Trials to Target Populations across Diverse Environments. In *International Conference on Machine Learning*. PMLR, 36170–36191.
[77] Baohong Li, Anpeng Wu, Ruoxuan Xiong, and Kun Kuang. 2024. Two-Stage Shadow Inclusion Estimation: An IV Approach for Causal Inference under Latent Confounding and Collider Bias. In *International Conference on Machine Learning*. PMLR, 28949–28964.
[78] Yunzhe Li, Kun Kuang, Bo Li, Peng Cui, Jianrong Tao, Hongxia Yang, and Fei Wu. 2020. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, Vol. 127. 4–22.
[79] Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 13201–13221.
[80] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. 2022. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*. PMLR, 13557–13603.
[81] Hao Liu, Yunze Li, Qinyu Cao, Guang Qiu, and Jiming Chen. 2019. Estimating Individual Advertising Effect in E-Commerce. *CoRR* abs/1903.04149 (2019).
[82] Michael J Lopez and Roee Gutman. 2017. Estimation of causal effects with multiple treatments: a review and new ideas*f*. *Statist. Sci.* (2017), 432–454.
[83] Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems*. 6446–6456.
[84] Bo Lu, Elaine Zanutto, Robert Hornik, and Paul R Rosenbaum. 2001. Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse. *J. Amer. Statist. Assoc.* 96, 456 (2001), 1245–1253.

[85] Jared Lunceford and Marie Davidian. 2004. Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in medicine* 23 (10 2004), 2937–60.
[86] Jing Ma. 2025. Causal inference with large language model: A survey. *Findings of the Association for Computational Linguistics: NAACL 2025* (2025), 5886–5898.
[87] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2021. Multi-Cause Effect Estimation with Disentangled Confounder Representation. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (2021).
[88] Brian D Marx and David Madigan. 1999. Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 94, 448 (1999), 467–477.
[89] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, and Krikamol Muandet. 2021. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction. In *International Conference on Machine Learning*, Vol. 139. 7512–7523.
[90] Roland A Matsouaka and Yunji Zhou. 2024. Causal inference in the absence of positivity: The role of overlap weights. *Biometrical Journal* 66, 4 (2024), 2300156.
[91] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105, 4 (2018), 987–993.
[92] Wang Miao, Wenjie Hu, Elizabeth L Ogburn, and Xiao-Hua Zhou. 2022. Identifying effects of multiple treatments in the presence of unmeasured confounding. *J. Amer. Statist. Assoc.* (2022), 1–15.
[93] Wang Miao and Eric Tchetgen. 2018. A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects. (08 2018).
[94] Abhirup Mondal, Anirban Majumder, and Vineet Chaoji. 2022. MEMENTO: Neural Model for Estimating Individual Treatment Effects for Multiple Treatments. In *International Conference on Information & Knowledge Management*. 3381–3390.
[95] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference.* Cambridge University Press.
[96] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. 2020. Dual Instrumental Variable Regression. In *Advances in Neural Information Processing Systems*.
[97] Preetam Nandy, Marloes H Maathuis, and Thomas S Richardson. 2017. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics* 45, 2 (2017), 647–674.
[98] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. 2021. VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments. In *International Conference on Learning Representations*.
[99] Xinkun Nie and Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (2021), 299–319.
[100] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. 2018. Learning independent causal mechanisms. In *International Conference on Machine Learning*. PMLR, 4036–4044.
[101] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. 2021. NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments. *CoRR* abs/2103.11175 (2021).
[102] Vern I. Paulsen and Mrinal Raghupathi. 2016. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.* Cambridge University Press.
[103] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[104] Judea Pearl. 2000. *Causality: Models, Reasoning and Inference.* Cambridge University Press.
[105] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).
[106] Jean Pouget-Abadie, Guillaume Saint-Jacques, Martin Saveski, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. 2019. Testing for arbitrary interference on experimentation platforms. *Biometrika* 106, 4 (2019), 929–940.
[107] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. 2021. Estimating Multi-cause Treatment Effects via Single-cause Perturbation. In *Advances in Neural Information Processing Systems*. 23754–23767.
[108] Rajesh Ranganath and Adler J. Perotte. 2018. Multiple Causal Inference with Latent Confounding. *CoRR* abs/1805.08273 (2018).
[109] Bo Ren, Xuefei Wu, Danielle Braun, Naveen Pillai, and Francesca Dominici. 2021. Bayesian modeling for exposure response curve via gaussian processes: Causal effects of exposure to air pollution on health outcomes. *arXiv preprint arXiv:2105.03454* (2021).
[110] James M. Robins and Sander Greenland. 1992. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* 3, 2 (1992), 143–155.
[111] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* (2000), 550–560.
[112] Paul R Rosenbaum. 1987. Model-based direct adjustment. *Journal of the American statistical Association* 82, 398 (1987), 387–394.
[113] 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.
[114] 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.

[115] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.
[116] Shiv Kumar Saini, Sunny Dhamnani, Aakash, Akil Arif Ibrahim, and Prithviraj Chavan. 2019. Multiple Treatment Effect Estimation Using Deep Generative Model with Task Embedding. In *The World Wide Web Conference*. 1601–1611.
[117] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. 2020. Cost-Effective and Stable Policy Optimization Algorithm for Uplift Modeling with Multiple Treatments. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 406–414.
[118] Pedro Sanchez and Sotirios A Tsaftaris. 2022. Diffusion Causal Models for Counterfactual Estimation. In *Conference on Causal Learning and Reasoning*. PMLR, 647–668.
[119] John D Sargan. 1958. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society* (1958), 393–415.
[120] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.
[121] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. 2020. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5612–5619.
[122] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.
[123] Hongwei Shang, Xiaolin Shi, and Bai Jiang. 2023. Network A/B Testing: Nonparametric Statistical Significance Test Based on Cluster-Level Permutation. *Journal of Data Science* 21, 3 (2023), 523–537.
[124] Claudia Shi, David M. Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. In *Advances in Neural Information Processing Systems*. 2503–2513.
[125] Xiao Shi, Weiming Miao, John C. Nelson, and Eric J. Tchetgen Tchetgen. 2020. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 2 (2020), 521–540.
[126] Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. 2020. A selective review of negative control methods in epidemiology. *Current epidemiology reports* 7 (2020), 190–202.
[127] Rahul Singh, Maneesh Sahani, and Arthur Gretton. 2019. Kernel Instrumental Variable Regression. In *Advances in Neural Information Processing Systems*. 4595–4607.
[128] Jeffrey A Smith and Petra E Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics* 125, 1-2 (2005), 305–353.
[129] Tamar Sofer, David B. Richardson, Elena Colicino, Joel Schwartz, and Eric J. Tchetgen Tchetgen. 2016. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science: a review journal of the Institute of Mathematical Statistics* 31, 3 (2016), 348–361.
[130] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. 1990. On the Application of Probability Theory to Agricultural Experiments. *Statist. Sci.* 5, 4 (1990), 465–472.
[131] Elizabeth A. Stuart. 2010. Matching Methods for Causal Inference: A Review and a Look Forward. *Statist. Sci.* 25, 1 (2010), 1–21.
[132] Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. 2025. Integrating Large Language Models in Causal Discovery: A Statistical Causal Approach. *Transactions on Machine Learning Research* (2025).
[133] Akira Tanimoto, Tomoya Sakai, Takashi Takenouchi, and Hisashi Kashima. 2021. Regret Minimization for Causal Inference on Large Treatment Space. In *International Conference on Artificial Intelligence and Statistics*, Vol. 130. 946–954.
[134] Henri Theil. 1953. *Repeated least squares applied to complete equation systems*. Central Planning Bureau.
[135] Elizabeth Tipton. 2013. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38, 3 (2013), 239–266.
[136] Hal R. Varian. 2016. Causal inference in economics and marketing. *Proc. Natl. Acad. Sci. USA* 113, 27 (2016), 7310–7315.
[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.
[138] Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. 2022. Generalization Bounds for Estimating Causal Effects of Continuous Treatments. In *Advances in Neural Information Processing Systems*.
[139] Yixin Wang and David M. Blei. 2018. The Blessings of Multiple Causes. *CoRR* abs/1805.06826 (2018).
[140] Yixin Wang and David M. Blei. 2021. A Proxy Variable View of Shared Confounding. In *International Conference on Machine Learning*, Vol. 139. 10697–10707.
[141] Yichao Wang, Huifeng Guo, Bo Chen, Weiwen Liu, Zhirong Liu, Qi Zhang, Zhicheng He, Hongkun Zheng, Weiwei Yao, Muyu Zhang, Zhenhua Dong, and Ruiming Tang. 2022. CausalInt: Causal Inspired Intervention for Multi-Scenario Recommendation. In *Conference on Knowledge Discovery and Data Mining*. 4090–4099.

[142] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics* 45, 10 (2013), 1113–1120.
[143] Raymond KW Wong and Kwun Chuen Gary Chan. 2018. Kernel-based covariate functional balancing for observational studies. *Biometrika* 105, 1 (2018), 199–213.
[144] P. G. Wright. 1922. *Tariff on animal and vegetable oils*. Macmillan.
[145] Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. 2022. Instrumental Variable Regression with Confounder Balancing. In *International Conference on Machine Learning*, Vol. 162. 24056–24075.
[146] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. 2022. Confounder Balancing for Instrumental Variable Regression with Latent Variable. *CoRR* abs/2211.10008 (2022).
[147] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, and Fei Wu. 2025. Instrumental variables in causal inference and machine learning: A survey. *Comput. Surveys* 57, 11 (2025), 1–36.
[148] Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319* (2024).
[149] Anpeng Wu, Haiyi Qiu, Zhengming Chen, Zijian Li, Ruoxuan Xiong, Fei Wu, and Kun Zhang. 2025. Causal Graph Transformer for Treatment Effect Estimation Under Unknown Interference. In *International Conference on Learning Representations*.
[150] Pengzhou Wu and Kenji Fukumizu. 2021. Intact-VAE: Estimating treatment effects under unobserved confounding. *arXiv preprint arXiv:2101.06662* (2021).
[151] Xuefei Wu, Danielle Braun, Joel Schwartz, Marianthi-Anna Kioumourtzoglou, and Francesca Dominici. 2020. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science Advances* 6, 29 (2020), eaba5692.
[152] Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, and Danielle Braun. 2022. Matching on generalized propensity scores with continuous exposures. *J. Amer. Statist. Assoc.* (2022), 1–29.
[153] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. 2021. Learning Deep Features in Instrumental Variable Regression. In *International Conference on Learning Representations*.
[154] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. 2021. Deep Proxy Causal Learning and its Application to Confounded Bandit Policy Evaluation. In *Advances in Neural Information Processing Systems*. 26264–26275.
[155] Xiaofang Yan, Younathan Abdia, Somnath Datta, KB Kulasekera, Beatrice Ugiliweneza, Maxwell Boakye, and Maiying Kong. 2019. Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in medicine* 38, 15 (2019), 2828–2846.
[156] 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 (2021), 74:1–74:46.
[157] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations*.
[158] Yi-Fan Zhang, Hanlin Zhang, Zachary C. Lipton, Li Erran Li, and Eric P. Xing. 2022. Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation. arXiv:2202.01336 [cs.LG]
[159] Qingyuan Zhao and Daniel Percival. 2016. Entropy Balancing is Doubly Robust. *Journal of Causal Inference* 5, 1 (2016).
[160] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift Modeling with Multiple Treatments and General Response Types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM.
[161] Ziyu Zhao, Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, Zhihua Wang, and Fei Wu. 2024. Networked Instrumental Variable for Treatment Effect Estimation with Unobserved Confounders. *IEEE Transactions on Knowledge and Data Engineering* (2024).
[162] Jiajing Zheng, Jiaxi Wu, Alexander D'Amour, and Alexander Franks. 2024. Sensitivity to unobserved confounding in studies with factor-structured outcomes. *J. Amer. Statist. Assoc.* 119, 547 (2024), 2026–2037.
[163] Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, and Liming Zhu. 2022. Learning to Infer Counterfactuals: Meta-Learning for Estimating Multiple Imbalanced Treatment Effects. *CoRR* abs/2208.06748 (2022).
[164] Peiwen Zhou, Billy K Chan, Yuk Kit Wan, Chun Ting Yuen, Gloria CY Choi, Xin Li, Cheuk Sum Tong, Shanshan Zhong, Jia Sun, Yi Bao, Sze Yan Mak, Man Z Chow, Joline V Khaw, Sing Yu Leung, Zhipeng Zheng, Lok W Cheung, Kuan Tan, Ka Hin Wong, Hing E Chan, and Aaron SC Wong. 2020. A three-way combinatorial crispr screen for analyzing interactions among druggable targets. *Cell Reports* 32, 6 (2020), 108020.
[165] Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, et al. 2024. Contrastive Balancing Representation Learning for Heterogeneous Dose-Response Curves Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17175–17183.
[166] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual Prediction for Bundle Treatment. In *Advances in Neural Information Processing Systems*.
[167] José R Zubizarreta, Caroline E Reinke, Rachel R Kelz, Jeffrey H Silber, and Paul R Rosenbaum. 2011. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician* 65, 4 (2011), 229–238.

Received 27 June 2024; revised 18 December 2025; accepted 10 January 2026