A Survey of Deep Causal Models and Their Industrial Applications

Zongyu Li[1,2], Zhenfeng Zhu[1,2*], Xiaobo Guo[3], Shuai Zheng[1,2], Zhenyu Guo[1,2], Siwei Qiang[3] and Yao Zhao[1,2]

[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China.
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, 100044, China.
[3]MYbank, Ant Group, Beijing, 100081, China.

*Corresponding author(s). E-mail(s): zhfzhu@bjtu.edu.cn; Contributing authors: zongyuli@bjtu.edu.cn; jefflittleguo.gxb@mybank.cn; zs1997@bjtu.edu.cn; zhyguo@bjtu.edu.cn; boyue.qsw@mybank.cn; yzhao@bjtu.edu.cn;

Abstract

The concept of causality plays a significant role in human cognition. In the past few decades, causal effect estimation has been well developed in many fields, such as computer science, medicine, economics, and other industrial applications. With the advancement of deep learning, it has been increasingly applied in causal effect estimation against counterfactual data. Typically, deep causal models map the characteristics of covariates to a representation space and then design various objective functions to estimate counterfactual data unbiasedly. Different from the existing surveys on causal models in machine learning, this paper mainly focuses on the overview of the deep causal models, and its core contributions are as follows: 1) we cast insight on a comprehensive overview of deep causal models from both timeline of development and method classification perspectives; 2) we outline some typical applications of causal effect estimation to industry; 3) we also endeavor to present a detailed categorization and analysis on relevant datasets, source codes and experiments.

Keywords: Deep causal models, causal effect estimation, causal learning, multiple treatment, continue dose treatment, covariates confounding learning

The authors have no relevant financial or non-financial interests to disclose.

1 Introduction

In general, causality refers to the connection between an effect and the cause of it. Causes and effects of this phenomenon are difficult to define, and we are often only aware of them intuitively[1]. Causal effect estimation is a process of drawing a conclusion about a causal connection based on the circumstances surrounding the occurrence of the effect and has a variety of applications in real-world scenarios[2]. For example, causal effect estimation of observational data in advertising[3–6], developing recommender systems that are highly correlated with causal treatment effect estimates[7–13], learning optimal treatment rules for patients in medicine[14–16], estimation of ITE in reinforcement learning[17–21], causal effect estimation tasks in natural language processing[22–26], emerging computer vision and language interaction tasks[27–31], education[32], and strategy resolutions[33–37], etc.

Deep learning largely contributes to the development of artificial intelligence when applied to big data[38–41]. In comparison with traditional machine learning, deep learning models are more computationally efficient, more accurate, and hold marvelous performance in various fields. However, many deep learning models are black boxes with poor interpretability since they are more interested in correlations than causality as inputs and outputs[42–44]. In recent years, deep learning models have been widely used for mining data for causality rather than correlation[33, 35]. Thus, deep causal models have become a core application for the issue of causal effect estimation in observation data with partial samples [16, 36, 37, 45]. At present, many works in the field of causal effect estimation utilize deep causal models to select reasonable treatment options[46–49].

With the emergence of big data, all trend variables are tended to be correlated[50], so discovering the latent causal relationships is becoming one open problem[51–53]. In terms of statistical theory, it is the most effective way to conduct **randomized controlled trials(RCT)** to infer causality. In other words, the sample is randomly assigned to a treatment or control group. Despite this, real-world RCT data are sparse and have several serious deficiencies. The studies involving RCTs require a large number of samples with little variation in characteristics, which is difficult to interpret and will further involve some ethical issues inevitably. As a matter of fact, it is also not wise to select subjects to try a drug or vaccine in the process of drug development. Therefore, causal effect estimation issues are commonly measured directly using observational data. A central question for obtaining counterfactual results is how to deal with observational data[54]. When observational data are analyzed, treatments are not randomly assigned and the performance of the post-treatment samples are significantly different from that of the general samples[33, 35]. Unfortunately, we can't observe alternative outcomes in theory since the counterfactual results[55] are not available to be observed. For a long time, most of researchers have mainly explored the use of the Potential Outcome Framework as a means of addressing the problem of causal effect estimation from observational data[56]. The Potential Outcome Framework is also known as the Rubin Causal Model[57]. In essence, causal effect estimation is closely connected to deep learning since it is conceptualized using Rubin Causal Model. In order to enhance the accuracy and unbiasedness of estimates, there are some works that attempt to combine deep networks and causal models. Just to illustrate several of them, e.g., the methods considering representations of distribution balance[33, 35, 36], exploiting the effects of covariates confounding learning[45, 58–60], the methods based on generative adversarial networks[37, 61–63] , and so forth[26, 49, 64].Due to the simultaneous development of deep learning and causal effect estimation, the problem of deep causal models has become more open and diverse[65–67].

In recent years, various perspectives have been discussed regarding causal effect estimation[1, 2, 68–76]. In Table 1, we list some of existing representative surveys and their highlight points. An in-depth analysis of the origin and development of causal effect estimation was provided in review[68], as well as the implications of causal learning for the development of causal effect estimation. Immediately after that, due to the rapid development of the field of machine learning, a detailed discussion of the relevance of graphical causal effect estimation to machine learning was presented in survey[69]. In addition, an overview of traditional and cutting-edge causal effect estimation methods and the comparison between machine learning and causal learning can be found in survey[1]. As one of hot focuses in recent years, the studies on the interpretability of machine learning have received great attention. An analytical summary of relevant interpretable artificial intelligence algorithms was elaborated in [70] by combining causal effect estimation with machine learning. Moreover, with the flourishing of causal representation learning, review[71] takes this new perspective to uncover high-level causal variables from low-level observations, strengthening the link between machine learning and causal effect estimation. In survey[75], the structural causal model of counterfactual intervention was comprehensively explained and summarized, and five classes of problems under causal machine learning are systematically analyzed and compared. Furthermore, in review[72], the authors discussed the way in which the latest progress of machine learning is applied to causal effect estimation, and provides a in-depth interpretation of how causal machine learning can contribute to the advancement in healthcare and precision medicine. As argued in review[73], causal discovery methods can be improved and sorted out based on deep learning, and can also be considered and explored from a variable paradigm perspective. Causal effect estimation in recommender systems is the focus of survey[74] , which explains how causal effect estimation can be used to extract causal relationships to enhance recommender systems. The Potential Outcome Framework in statistics has long served as a bridge between causal effect estimation and deep learning. As a starting point, review[2] analyzes and compares different classes of traditional statistical algorithms and machine learning that satisfy these assumptions. Recently, survey[76]

Table 1 Highlights of existing surveys on causal learning

Surveys	Highlights
The Development of Causal Reasoning[68]	The origin and development of causal reasoning
Causality for Machine Learning[69]	The connection between graphical causal effect estimation and machine learning
Causal Inference[70]	Machine Learning Interpretability for Counterfactual Causal Reasoning
Toward Causal Representation Learning[71]	Exploring causal variables in data through causal representation learning
Causal Machine Learning for Healthcare and Precision Medicine[72]	Causal Machine Learning in clinical decision support systems
A Survey of Learning Causality with Data: Problems and Methods[1]	The relationship between causal learning and machine learning in big data
A Survey on Causal Inference[2]	Causal effect estimation of observational data in the potential outcomes framework
A Review and Roadmap of Deep Learning Causal Discovery in Different Variable Paradigms[73]	An improved exploration of causal discovery from the perspective of deep learning and variable paradigms
Causal Machine Learning: A Survey and Open Problems[75]	Five types of causal machine learning problems under structural causal models
Causal Inference in Recommender Systems: A Survey and Future Directions[74]	Optimize the recommendation system by extracting causality through causal reasoning
Deep Causal Learning: Representation, Discovery and Inference[76]	The significance of deep learning for solving traditional causal learning problems
A Survey of Deep Causal Models and Their Industrial Applications	Causal effect estimation in deep models with biased sample observations and challenges in current industrial applications

provides an overview of causal learning from Structural Causal Model, Potential Outcome Model, Deep Neural Network and Deep Causal Discovery, explores the internal mechanism of deep learning to improve the significance of causal learning. Different from the above review, we penetrate into the territory of biased sample observation of causal effect estimation, provide a systematic and comprehensive overview of the deep model in representation learning, debias estimation, counterfactual inference and other perspectives. It is worth mentioning that we provide a comprehensive conclusion of the latest applications of deep causal models in industry. This survey makes three core contributions: 1) we cast insight on a comprehensive overview of deep causal models from both temporal development and method classification perspectives; 2) we outline some typical applications of causal effect estimation to industry; 3) we also try to present a detailed categorization and analysis on relevant datasets, source codes and experiments.

The rest of the paper is outlined as follows. In Section 4, the deep causal model is comprehensively elaborated. Next, we categorize the deep causal modeling methods into five groups in Section 5, including learning balanced representation, covariates confounding learning, GANs-based counterfactual simulation, time series causal estimation, and multi-treatment and continuous dose treatment models. Additionally, in Section 6, we summarize the industrial applications of causal reference. After that, the relevant experimental guidelines are listed in Section 7. Finally, we conclude this article in Section 8.

2 Preliminaries

In this section, we present the background knowledge of causal effect estimation, including task descriptions, mathematical concepts, and pertinent assumptions.

Basically, the aim of causal effect estimation is to estimate the change in outcome that will occur if a different treatment are implemented[2]. Imagine that there are several treatment plans A, B, C, and so on, all of which have different cure rates, and the change in the cure rate is the result of the treatment scheme. Realistically, we can't apply different treatment regimens to the same group at the same time. As opposed to RCT, the main problem to be solved in observational research is the lack of counterfactual data. In other words, the challenge we face is how to find the most effective treatment plan based on past experimental diagnosis and medical history of the patient.

2.1 Definitions

Here, to make the survey self-consistent, we first give some basic definitions under the Potential Outcome Framework[57]. In particular, causality is defined as the result of a treatment scheme applied to a sample, which can either be a specific behavior, a specific method, or some specific treatment scheme. The followings are the concepts related to causal effect estimation, which are benchmarked against the relevant basic definitions in the survey[2].

Definition 1 *Treatment Effect Estimation(Causal Effect Estimation):*Estimate the change in outcome of a given sample receiving an intervention.

Treatment effect estimation and lift modelling have the same goal, the difference is: the lift model is on random experimental data, treatment effect estimates usually need to adhere to the necessary assumptions, applied to experimental and observational data.

Definition 2 *Treatment:* A treatment refers to a scheme or action applied to a sample.

As a medical term, a drug scheme is a treatment. For binary treatments, $T = 1$ is the *treated group*, and $T = 0$ is the *control group*. Multiple treatment can be indicated by the $T \in \{0, 1, 2, \ldots, T_N\}$, where $N + 1$ denotes the total number of treatments.

Definition 3 *Observed outcome:* An observational outcome, also known as a factual outcome, is a measure of how the sample's outcomes applied to the treatment.

In the case of a specific treatment, observed outcomes can be displayed in Y^F, where $Y^F = Y(T = T_i)$. A donation of in the amount of Y_{T_i} is made as *potential outcome*.

Definition 4 *Counterfactual outcome:* A counterfactual outcome is the outcome that differ from a factual outcome.

With binary treatments, counterfactual outcome is denoted as Y^{CF}, and $Y^{CF} = Y(T = 1 - T_i)$. Assuming multiple treatment, let $Y^{CF}(T = T_i^{'})$ donate the counterfactual result of treatment $T_i^{'}$.

Definition 5 *Dose:*A dose refers to the amount of drug taken continuously during a particular course of treatment.

In general, there are many medical treatments involving continuous dose parameters, such as vasopressors. A set of continuous dose regimens can be donated as D_T, the treatment-specific factual doses can be donated as can be donated as D^F, and $D^F = D(T = T_i)$. Simultaneously, counterfactual dose can be donated as $D^{CF}(T = T_i^{'})$.

Definition 6 *Dose-response curve:*A dose-response curve indicates the effect of the samples' response over time after receiving different doses of the intervention.

A better fit to the dose-response curve could make the model more robust and expressive in continuous dose treatments. The actual and counterfactual outcomes on the dose-response curves can be expressed as the sets $Y^F(D^F, T_i)$ and $Y^{CF}(D^{CF}, T_i^{'})$.

Definition 7 *Covariates:* Covariates are variables that are unaffected by treatment choice.

Generally, covariates in the healthcare setting refer to the patient's demographic, medical history, experimental data, etc., usually denoted by X. Covariates can be separated into confounding and non-confounding variables, specifically divided into three categories[58]: instrumental factor I, which only affects treatment T; confounding factor C, which contributes to both treatment T and outcome Y; and adjustment factor A, which only determines outcome Y.

2.2 Assumptions

After understanding the basic definition of causal effect estimation, the following three assumptions, which are derived from papers[2, 70], are commonly required to achieve the estimation of causal treatment effects.

Assumption 1 Stable Sample Treatment Value (SSTV): One sample's response to treatment is independent of the assignment of other samples.

Based on this assumption, there is no interaction between samples, and there is only one version of each treatment option. SSTV assumption can be expressed as $P(Y_i \mid T_i, T_i^{'}, X_i) = P(Y_i \mid T_i, X_i)$.

Assumption 2 Ignorability: Given the covariate X, the treatment distribution T is independent of the potential outcomes.

On the assumption of ignorability, there should be no unobserved confounding factors. In other words, $T \perp\!\!\!\perp Y(T = T_i), Y(T = T_i^{'}) \mid X$ needs to be satisfied.

Assumption 3 Overlap: When given the observed variables, each sample has nonzero probability to receive either treatment status.

In order to estimate the counterfactual treatment effect, it must be assumed that each sample can implement any treatment option, otherwise the overlap assumption will not be valid. That is, $0 < P(T = T_i \mid X = x) < 1$ and $0 < P(T = T_i^{'} \mid X = x) < 1$

3 Treatments and Metrics

This section provides an analysis and description of the different performance metrics adopted for different classical application scenarios. In addition to the basic metrics in survey[2], we expand the evaluation from binary to multiple and continuous dose cases.

3.1 Binary treatment

Treatment refers to the action of a sample or a subject. In the medical field, a medication regimen for a patient is a treatment. When there are only two treatment options, the group of units applied with treatment $T = 1$ is the treated group, and the group of units with $T = 0$ is the control group, which can be referred to as the binary treatment[2].

In the binary treatment situation, the most basic and common performance metric is **Average Treatment Effect(ATE)**, which is defined as[77]:

$$\mathrm{ATE} = \mathbb{E}[Y(T = 1) - Y(T = 0)], \tag{1}$$

where $Y(T = 1)$ and $Y(T = 0)$ indicate the results of the treatment and control groups in the population. This metric is commonly used to estimate causal effects for the well-known dataset IHDP[78]. Differentiate between treatment and control groups by whether intensive high-quality child care is applied, and the outcome is a score on the cognitive test for the infant.

In a sample set, the treatment effect is called the **Conditional Average Treatment Effect (CATE)** that is given as follows[2]:

$$\mathrm{CATE} = \mathbb{E}[Y(T = 1) \mid X = n] - \mathbb{E}[Y(T = 0) \mid X = n], \tag{2}$$

where $Y(T = 1) \mid X = n$ and $Y(T = 0) \mid X = n$ represent the results under the sample set for the treatment and control groups with $X = n$, respectively. Since different treatments have different effects on different example sets, CATE is also known as heterogeneous treatment effect. The treatment effect at the individual level is called **Individual Treatment Effect (ITE)**, which is defined as[2]:

$$\mathrm{ITE}_n = Y_n(T = 1) - Y_n(T = 0), \tag{3}$$

where $Y_n(T = 1)$ and $Y_n(T = 0)$ represent the results of the sample in the treatment and control groups.

Another evaluation metric to consider is PEHE (Precision in Estimation of Heterogeneous Effects) It is also helpful to keep in mind another evaluation metric, which is called **Precision in Estimation of Heterogeneous(PEHE)**. Regardless of fact or counterfactual outcomes, PEHE, as defined below[37], can perform unbiased estimates:

$$\text{PEHE} = \frac{1}{N}\sum_{n=1}^{N}\left(Y_1^F(n) - Y_0^F(n) - \left(Y_1^{CF}(n) - Y_0^{CF}(n)\right)\right)^2 \tag{4}$$

where $Y_1^F(n)$, $Y_0^F(n)$ and $Y_1^{CF}(n)$, $Y_0^{CF}(n)$ respectively indicate unbiased estimates of the facts and the counterfactuals for the treatment and control groups. The Twins[45] dataset is typical of the application of this metric, and the result is a one-year mortality rate for children.

For the treated group, the treatment effect is referred to as the **Average Treatment effect on the Treated group (ATT)**, which is defined as[2]:

$$\text{ATT} = \mathbb{E}[Y(T=1) \mid T=1] - \mathbb{E}[Y(T=0) \mid T=1], \tag{5}$$

where $Y(T=1) \mid T=1$ and $Y(T=0) \mid T=1$ correspond to potential treated and control outcome of the treated group respectively. This metric is available for the jobs dataset[79], treatment group take part in vocational training, while the control group are not and the outcome is employment status.

Since only factual data is available for Jobs dataset, the testing set is from RCT. Therefore, heterogeneity effect indicators only can be described by **Policy Risk** ($\mathcal{R}_{pol}(\pi)$), which is defined as[35]:

$$R_{\text{pol}}(\pi) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left[1 - \left(\sum_{i=1}^{K}\left[\frac{1}{|\Pi_i \cap T_i \cap E|}\sum_{X(n)\in\Pi_i\cap T_i\cap E} Y_i^F(n) \times \frac{|\Pi_i \cap E|}{|E|}\right]\right)\right] \tag{6}$$

where $\Pi_i = \{X(n) : i = \arg\max Y^{\hat{C}F}\}, T_i = \{X(n) : t_i(n) = 1\}$, and E is the subset of RCT.

3.2 Multiple treatment

Unlike the binary treatment problem, the multiple treatment problem can be described by the following stereotype[34]: For each sample, the potential outcomes are represented as a vector Y with N entries ${Y_j}^F$ where each entry corresponds to the outcome when applying one treatment T_j out of the set of N available treatments $T = \{0, 1, 2, \ldots, T_N\}$, with $j \in \{0, 1, 2, \ldots, N-1\}$. The set of available treatments can contain more than two treatments. As training data, samples k and their observed factual outcomes Y^F are received when applying one treatment T_i , the other outcomes can not be observed. The purpose is to train a predictive model $\hat{f}$ that is able to estimate the entire potential outcomes vector $\hat{Y}$ with k entries Y_j.

With the exception of ATE and CATE, an accurate estimate of treatment effect can be determined by using **Root Mean Square Error (RMSE)** for all subgroups, which is defined as[80]:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\mid T\mid}\sum_{j\in T}\left(Y^F(n,j) - Y^{CF}(n,j)\right)^2} \tag{7}$$

where T is the set of all possible combinations of treatments, N is the number of observations. $Y^F(n,j)$ displays the true outcome when the j^{th} subset is applied to compute the n^{th} observation, whereas $Y^{CF}(n,j)$ displays the predicted outcome when the j^{th} subset is applied to compute the n^{th} observation. The **Absolute Error** can be calculated in the same way. This metric is commonly used for the News[34] dataset, which the treatments are the choice of viewing tools, such as smartphones, tablets, desktops, and TVs.

To measure the average of the RMSE between the actual and estimated difference between ITE for each treatment with no treatment ITE, We can incorporate multiple treatment into one calculation criterion, which is called **Average PEHE**[80]:

$$\text{AveragePEHE}_j = \frac{1}{|T|}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left((Y^F(n,j) - Y^F(n,0)) - (Y^{CF}(n,j) - Y^{CF}(n,0))\right)^2}, j \in (T - T_0) \tag{8}$$

where T_0 represents the sample set with no treatment applied. It is worth mentioning that the latest research on estimating the effect of multi-causal treatment COVID-19[81, 82]. Average PEHE typically serve as performance metrics in dataset CHESS.

3.3 Continuous dose treatment

Vary from binary and multivariate treatment issues, the continuous-dose treatment issue can be formulated as follows[62]: Consider to receive observations of the form $(\mathbf{x}^i, T_f^i, Y_f^i)$ for $i = 1, ..., N$, where, for each i, these are independent realizations of the random variables $(\mathbf{X}, T_f, Y_f)$, refer to $\mathbf{X}$ as the feature vector lying in some feature space $\mathcal{X}$, containing pre-treatment covariates. The treatment random variable, T_f, is in fact a pair of values $T_f = (W_f, D_f)$ where $W_f \in \mathcal{W}$ corresponds to the *type* of treatment being administered which lies in the discrete space of N treatments, $\mathcal{W} = \{w_1, ..., w_k\}$, and D_f corresponds to the *dosage* of the treatment, which, for a given treatment w lies in the corresponding treatment's dosage space, $\mathcal{D}_w$ (e.g. the interval $[0, 1]$). The set of all treatment-dosage pairs can be defined as $\mathcal{T} = \{(w, d) : w \in \mathcal{W}, d \in \mathcal{D}_w\}$.

In continuous dose treatment situation, the sample dose-response curve6 is adopted as the metric. Furthermore, the metric on the test set is different[48]. Therefore, we can use **Mean Integral Squared Error (MISE)** to measure the accuracy of the estimation of the patient treatment effect on the dose space, which is defined as[48].

$$\text{MISE} = \sqrt{\frac{1}{K}\frac{1}{N}\sum_{T_n \in \mathcal{T}}\sum_{n=1}^{N}\int_{\mathcal{D}_{T_n}}\left(Y_n(T_n,u) - \hat{Y}_n(T_n,u)\right)^2 \, \mathrm{d}u} \tag{9}$$

where T is the set of treatments in the space, K is the number of samples, and u is the treatment dose taken. For a given treatment, T_n lies within the dose space D_{T_n}. Moreover, $Y_n(T_n, u)$ and $\hat{Y}_n(T_n, u)$ denote the results predicted by the model and the results at the optimal treatment dose, respectively.

In addition, the **Mean Dose Policy Error (DPE)** is another marvelous metric of a model's ability to predict the optimal dose point for each individual treatment, which can be defined by [48]:

$$\text{DPE} = \sqrt{\frac{1}{K}\frac{1}{N}\sum_{T_n \in \mathcal{T}}\sum_{n=1}^{N}\left(Y_n\left(T_n, D^*_{T_n}\right) - Y_n\left(T_n, \hat{D}^*_{T_n}\right)\right)^2} \tag{10}$$

where, $D^*_{T_n}$ and $\hat{D}^*_{T_n}$ represent the true optimal dose and the model-determined optimal dose under a treatment, respectively. With **Sequential Least Squares Estimation**, the optimal dose point for the model can be determined.

In order to compare the optimal treatment dose pair selected by the model with the true optimal treatment dose pair, the mean **policy error (PE)** is given as[48]:

$$\text{PE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(Y_n\left(T^*_n, D^*_{T_n}\right) - Y_n\left(\hat{T^*_n}, \hat{D}^*_{T_n}\right)\right)^2} \tag{11}$$

where T^*_n and $\hat{T^*_n}$ represent the optimal treatment and the optimal treatment determined by the model, respectively. By calculating the optimal dose for each treatment and then selecting the treatment that yields that optimal dose, the optimal dose pair for the model can be selected. The above metrics are typically used in the TCGA[83] dataset. Among them, drug therapy, chemotherapy, and surgery are the treatment options, and the outcome is a risk of cancer recurrence after treatment.

Additionally, many researchers have flexibly applied simulation datasets to differ scenarios. In order to provide more ablation experiments and prove the robustness of the model. A detailed description of the relevant datasets can be found in Section 7.

4 Development of Deep Causal Models

In recent years, there is growing dissatisfaction within industry with current machine learning algorithms, 87% of machine learning projects never make it beyond an experimental phase into production, according to Forbes. However, in the context of big data, with the development of deep learning, the decision-making ability of causal effect estimation has been significantly improved. A recent global survey found that 83% of Ctos and senior data scientists agree that causal deep learning can lead to a more robust business environment model, and 60% of senior data scientists intend to make a significant investment in deep causal models.

With a solid understanding of the basic definitions and model measures of causal effect estimation, this section moves to the core of the paper. We provide an overview of deep causal models during the last several years and a detailed classification of them.

4.1 A timeline of development

In the past few years, research on deep causal models has advanced considerably, and they have become more accurate and efficient in estimating causal effects. In Figure1, we present the development timeline about 50 classical deep causal models from June 2016 to November 2022. It also displays categories of models and the relationship between essays.

Deep causal models have emerged since 2016. For the first time, Johansson et al. publish **Learning Representations for Counterfactual Inference**[33], and propose the algorithm framework BNN and BLR[33], in which deep learning is combined with the causal effect estimation, and the causal effect estimation problem is transformed into a domain adaptation problem. Since then, a number of models, including DCN-PD[84], TARNet and CFRNet[35], have been proposed since then. It is important to note that the CEVAE[45] model proposed by Louizos et al. in December 2017, which is based on classical structural Variational Autoencoders, focuses on the impact of confounding factors on the estimation of causal effects.

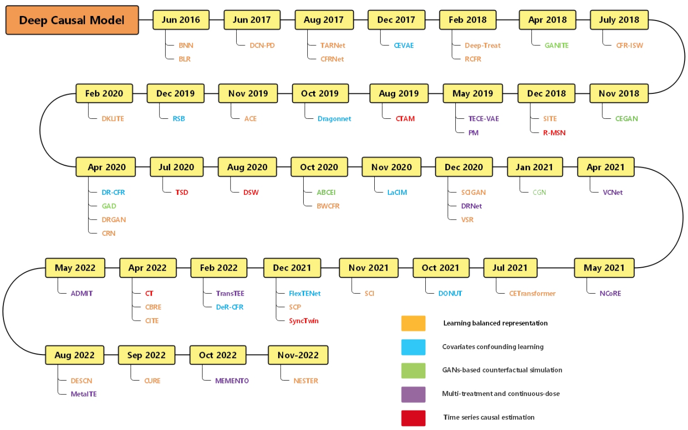

Fig. 1 Overview of development timeline of classical deep causal models

In 2018, and going forward into 2019, there emerged an increasing interest in causal representation learning with representative works including DeepTreat[16] and RCFR[85] models. After the launch of the GANITE[37] model, the use of generative adversarial model architecture for counterfactual estimation becomes mainstream in the field of causal effect estimation. In accordance with the previous works, some novel optimization ideas were proposed in CFR-ISW[86], CEGAN[61], SITE[36].

By applying recurrent neural networks, the R-MSN[64] model aims to solve the problem of continuous dose of multi-treatment time series, which opened up a new theory of deep causal models. Furthermore, PM[34] and TECE-VAE[80], proposed in 2019, attempt to address the issue of estimating causal effects associated with multiple discrete treatment. As a follow-up, the CTAM[26] begins to focus on estimating causal effects for textual data; the Dragonnet[59] adhibits regularizations and propensity score networks into causal models for the first time; the ACE[46] attempts to extract fine-grained similarity information from representation space. After that, RSB[87] model adopts deep representation learning network and PCC regularization for covariate decomposition, uses instrumental variables to control selection bias, and uses confounding factors and moderating factors to predict results.

In 2020, deep causal models got a rapid boom. For DKLITE[47] model, it effectively combines the deep kernel trick and posterior variance regularization. Then, DR-CFR[88] was proposed to decouple selection bias for covariates; GAD[89] focuses on the causal effect of continuous dose treatment; DRGAN[90] defines an innovative generative adversarial network for fitting sample dose effect curves; and CRN[91] estimates time-varying treatment effects by combining counterfactual recurrent neural networks. Following estimating time series causal effects under multi-cause confounding, TSD[92] turns to the estimation of time series causal effects. In the aspect of learning the latent representation space, ABCEI[93] achieves a marvelous balance between the covariate distribution of treatment and the control groups by using GAN. As two variants of previous works[35, 45], both BWCFR[94] and LaCIM[95] are further optimized for the model structure. Additionally, in 2020, SCIGAN[62] and DRNet[48] extended the task of continuous dose treatment to arbitrary quantity, while VSR[96] aggregated the deep latent variables in a reweighted manner.

Up to now since 2021, deep causal models have become more innovative, open, and flexible. The VCNet[49] model implemented an estimator of continuous mean dose-response curves. Moreover, CGN[63] proposed more robust and interpretable classifiers that explicitly expose the causal structure of tasks. As for NCoRE[97], it used cross-treatment interaction models to explain the process of multi-treatment combination at the potential causal level. After that, inspired by the success application of Transformer in many fields, CETransformer[98] was proposed to characterize the covariates by focusing the attention on the correlation between the covariates. In addition, DONUT[99] and DeR-CFR[58] are optimized for the two previous works[59, 88] respectively. In [65], the subspace methods was further explored for causal representation learning, which formulate the SCI model. From the perspective of multi-task learning e, a multi-task adaptive learning architecture was proposed by FlexTENet[100], in which CATE estimation architecture adaptively learns what to share between the PO functions. Aside from that, SCP[101] attempted to estimate the multi-factorial treatment effects using a two-step procedure. To construct synthetic twin matching representation, SyncTwin[102] utilized the temporal structure in the potential outcomes.

In 2022, TransTEE[66] extended the distribution balance approach for spatial representations to continuous, structured, and dose-dependent treatments, applying the latest theory of deep learning to expand the depth and breadth of causal effect estimation. Furthermore, DESCN[103] obtaind comprehensive information on treatment propensity, response and hidden treatment effects through a crossover network by means of a multi-task learning. In addition, CBRE[104] employed a cyclic structure to maintain the original data attributes and minimize information loss during the transformation of data into a latent representation space. Immediately after, CITE[105] leveraged self-supervised information embedded in the data and achieves a balanced and predictive representation by appropriately utilizing causal prior knowledge. Moreover, ADMIT[106] provided a generalized bound for estimating ADRF, alleviating selection bias, and offering a discrete approximation of IPM distance with theoretical guarantees. Beside that, CT[107] combined Transformer with the LSTM network to capture complex long-term dependencies between time-varying confounders. Until recently, CURE[108] pre-trained large-scale unlabeled patient data to learn representative background patient representations, and then fine-tunes the labeled patient data for treatment effect estimation. In particular, NESTER[109] utilized neural symbolic programming to address treatment effect estimation problems, allowing for the implicit encoding of inductive bias within a domain-specific language.

4.2 Model classification

Through sorting out the existing models, we can find that the current deep causal models are mainly studied from the following aspects that include : 1) Learning balanced representations; 2) Covariate confounding learning; 3) Time series causal learning; 4) GANs based Counterfactual simulation; and 5) Multi-treatment and continuous dose treatment. In Figure 2, we present a detailed classification of the current deep causal model.

- **Learning balanced representations:** This type of approach has long been a popular research. The core ideology is to use the encoder to map the covariates X to the representation space Φ, combine the processing T, adopt the network h to predict the output outcome Y, and minimize the distribution distance $\mathrm{disc}_{\mathcal{H}}$ between the factual and counterfactual outcomes. Classical architectures include BNN[33], CFRNet[35], SITE[36], ACE[46], DKLITE[47], SCI[65], etc.

- **Covariate confounding learning:** This type of approach aims to decomposition of covariant relation in theory. Its main application schemes are unbiased estimation of covariates and removal of confounding factors using decoupling, reweighting, codec reconstruction, etc. The typical structures include CEVAE[45], Dragonnet[59], DeR-CFR[58], LaCIM[95], DONUT[99], FlexTENet[100], and so on.
- **Time series causal estimation:** Temporal causal estimation has been widely concerned. Using RNNs to track contextual covariate information and handle time-varying confounding bias is a long-standing solution adopted by many models. Some typical architectures are R-MSE[64], CTAM[26], CRN[91], TSD[92], and so on.
- **GANs-based counterfactual simulation:** With the great success of GANs in data synthesis in recent years, it is also widely adopted to solve causal effect estimation problems. Two schemes are usually involved in Using GAN networks for counterfactual simulation, i.e., generating the counterfactual output outcomes or balanced representation space distributions. Such classical frameworks include GANITE[37], CEGAN[61], ABCEI[93], CETtrnaformer[98], etc.
- **Multi-treatment and continuous dose treatment:** The issues of Multiple treatment and continuous dose treatment are one of recent research hotspots in deep causal learning. In general, such issues can be further simplified and structured using schemes such as matching, variational autoencoders, hierarchical discriminators, and multi-headed attention mechanisms. The classical models include PM[34], TECE-VAE[80], DRNet[48], SCIGAN[62], VCNet[49], TransTEE[66], etc.

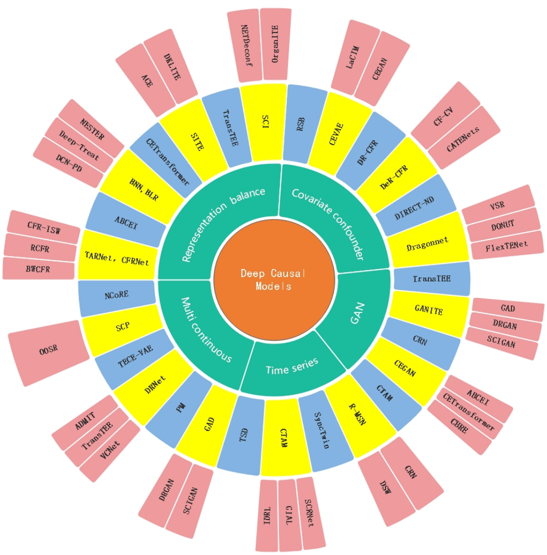

Fig. 2 Taxonomy of deep causal models

It is worth pointing out, as different application scenarios arise, new research focuses and methods in the aspect of deep causality model will also emerge continuously. In the following section, we will give a detailed introductions to the typical models according to the taxonomy of deep causal models motioned above.

5 Typical Deep Causal Models

Deep learning is being utilized more often in various fields such as healthcare, education, and economy to draw causal relationships from counterfactual data. Unlike conventional deep causal models that map covariates to a representation space, the objective function can provide a more impartial estimation of counterfactual data. Unlike the brief overview of the various classical models that are categorized from the different research perspectives for deep causal models, Table 2 summarizes the classical network architecture applied by those typical deep causal models. In addition, in the following detailed descriptions on the typical deep learning-based causal models, the issues and challenges that these models face are also discussed.

Table 2 Highlights of deep frameworks on classical causal models

Models	GAN	AE	RNN	Transformer
BNN[33]		✓		
CFRNet[35]		✓		
CEVAE[45]		✓		
GANITE[37]	✓			
CEGAN[61]	✓			
SITE[36]		✓		
R-MSN[64]			✓	
TECE-VAE[80]		✓		
CTAM[26]	✓	✓		
Dragonnet[59]		✓		
ACE[46]		✓		
DKLITE[47]		✓		
GAD[89]	✓			
CRN[91]	✓		✓	
TSD[92]			✓	
ABCEI[93]	✓	✓		
SCIGAN[62]	✓	✓		
DRNet[48]		✓		
VCNet[49]		✓		
CETransformer[98]	✓	✓		✓
DeR-CFR[58]		✓		
SCI[65]		✓		
SCP[101]		✓		
CGN[63]	✓	✓		
SyncTwin[102]			✓	
TransTEE[66]	✓	✓		✓
CURE[108]		✓		✓
DSW[110]		✓	✓	
CBRE[104]	✓	✓		
CT[107]	✓	✓	✓	✓

5.1 Learning balanced representation

In reality, test and training data often have related but not identical distributions, which is a challenge for statistical learning theories. To solve this problem in the field of causal effect estimation, deep learning models that learn causality instead of correlation are used. Unlike in RCTs, there is no standard treatment assignment strategy for observational data, and selection bias caused by known and unknown covariates can lead to different fact and counterfactual distributions. Therefore, to predict counterfactual outcomes by learning from factual data, causal effect estimation needs to be transformed into a domain adaptation problem.

For counterfactual results to be predicted, effective feature representations are necessary, especially the balanced distributions. According to Johansson et al, BNN[33] is an algorithmic framework as shown in Figure 3 for counterfactual reasoning that transforms the causal effect estimation problem into a representation distribution balance problem.

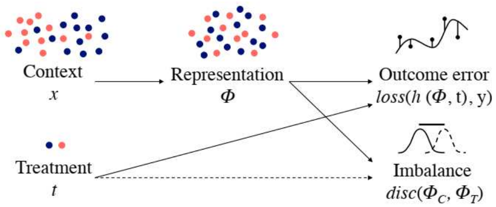

Fig. 3 Balancing Neural Network (BNN) [33]: Counterfactual inference networks based on representation distribution balance

Upon mapping the covariates to the representation space, the encoder makes use of a two-layer fully connected neural network, balances the distribution distance of the representation space, and then derives the counterfactual results using another two-layer fully connected network. The used regression function is as follows:

$$B_{\mathcal{H},\alpha,\gamma}(\Phi,h)=\frac{1}{n}\sum_{i=1}^{n}\mid h\left(\Phi\left(x_i\right),t_i\right)-y_i^F\mid+\alpha\,\mathrm{disc}_{\mathcal{H}}\left(\hat{P}_{\Phi}^{F},\hat{P}_{\Phi}^{CF}\right)+\frac{\gamma}{n}\sum_{i=1}^{n}\mid h\left(\Phi\left(x_i\right),1-t_i\right)-y_{j(i)}^F\mid \tag{12}$$

where the encoder network is denoted by Φ, a predictor network by h, and the metric function is $\mathrm{disc}_{\mathcal{H}}$ that represents the distance between the two distributions. Aside that, Eq.(12) minimizes the error of the training set facts.

As an innovative method for measuring the spatial distribution distance between treatment groups and control groups, the literature[35] proposed a CFRNet network structure based on BNN[33] and adopted MMD and WASS for spatial distribution distance representations. When the network is trained, the imbalance penalty is calculated based on the explicit boundary of the distance, and the loss is calculated separately for the treatment group and the control group. As well as adding multiple neural network layers between results predition layer, DCN-PD[84] combined multi-task deep neural networks with propensity score dropout.

On the basis of the CFRNet[35] model, both RCFR[85] and CFR-ISW[86] used the Propensity score to re-weight the representative spatial feature region and the sampling objective function; Atan et al. proposed an unbiased autoencoder network Deep-Treat[16] framework, applying a feedforward neural network to learn the optimal treatment strategy. While it reduces the loss of representation reconstruction as well as the information loss in space, the selection bias is also narrowed.

To maintain local similarity and balance of data representing the treatment and control groups and to improve individual treatment outcomes. Yao et al. proposed the SITE[36] method, which combines position-dependent depth metric PDDM with midpoint distance minimization MPDM into the representation space, and predicts the potential outcomes using a binary result network. In this case, the following loss function is used:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FL}} + \beta \mathcal{L}_{\mathrm{PDDM}} + \gamma \mathcal{L}_{\mathrm{MPDM}} + \lambda \|W\|_2 \tag{13}$$

it employs various techniques to minimize errors and prevent overfitting, such as $\mathcal{L}_{\mathrm{FL}}$ for comparing predicted and actual factual results, $\mathcal{L}_{\mathrm{PDDM}}$ and $\mathcal{L}_{\mathrm{PDDM}}$ for evaluating the MPDM and PDDM, and L_2 regularization for the parameter W.

According to SITE[36], ACE[46] proposed a balanced and adaptive similarity regularization structure to extract spatially fine-grained similarity information; in [47] DKLITE was proposed by applying the deep kernel regression and a posterior regularization to learn the spatial domain overlap information; and BWCFR[94] re-weighted the spatial feature distribution for the domain overlap region. In addintion, CBRE[104] utilized cyclically balanced representation learning to retain crucial information about the original data attributes and minimize information loss when converting data into a latent representation space during counterfactual inference; CITE[105] constructed a framework for comparing individual treatment effects by leveraging self-supervised information hidden in the data, achieving balanced and predictive representations while appropriately utilizing causal prior knowledge.

Unlike the above models, many other works attempt to combine representation distribution balancing with other deep network models. By combining GANs with a mutual information estimator regularization structure, ABCEI[93] tries to balance the covariate distributions of the treatment and control groups in the representation space; In [98], CETransformer was proposed to apply the attention mechanism to focus on the relationships between covariates and then learn a balanced representation distribution; As TransTEE[66] extends the balanced representation distribution method to Continuous, Structured, and Dose-Related treatments, it makes causal effect estimation a more open-ended problem; CURE[108] designed a new sequence encoding for longitudinal (or structured) patient data and merge structure and time into patient embeddings; DESCN[103] learned the processing and response functions jointly across the sample space to avoid processing bias, and used an intermediate pseudo-processing effect prediction network to mitigate sample imbalance; In [65], SCI inducted the concept of a subspace as shown in Figure 4, integrating the covariates into a common subspace, a treatment subspace, and a control subspace simultaneously, thereby obtaining a common representation and two specific representations. Afterwards, the common representation is connected to the specific representation of the treatment group and of the control group, and two potential results are obtained from the reconstruction and prediction network. Based on SCI, NETDECONF[111] used network structure information to infer hidden confounding in the observational data. A personalized treatment effect model that allocates treatments according to scarcity and estimates potential outcomes is proposed by OrganITE[112]. It is worth mentioning that NESTER[109] employed inductive bias and heuristics to design multi-head neural network structures and regularizers, and applied neural symbolic programming instead of traditional methods for treatment effect estimation.

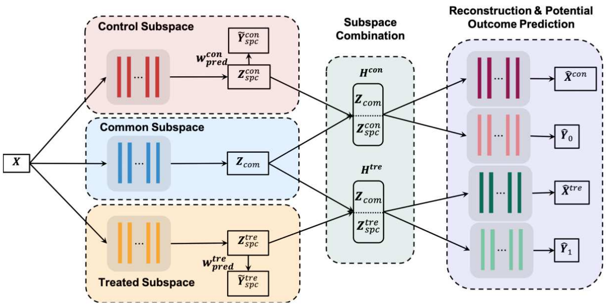

Fig. 4 Subspace Learning Based Counterfactual Inference network (SCI) [65]: Estimating individual causal effects while preserving information about specific and common subspaces

Due to the improvement in the feasibility of estimating causal effect, the representation distribution balance is becoming the mainstream way, but it is still limited to estimating individual treatment effects, and thus is hard to expand to a broader range of applications like multi-treatments and continuous dose treatments.

5.2 Covariates confounding learning

The main issue in causal effect estimation is estimating the treatment effect when given a covariate, a treatment, and a predicted outcome. By identifying and correcting for confounders, it is possible to estimate causal effects with greater accuracy from observational data. Nevertheless, in practical cases, there are potential confounders of noise and uncertainty, as well as some non-confounders. For this reason, mining potential confounders and decoupling the associated covariates are important methods to learn unbiased representations of counterfactuals from observed data.

A CEVAE[45] model structure was first proposed by Louizos et al. to capture hidden confounding factors with VAEs in the presence of noise and uncertain confounding factors, and perform treatment and prediction. On the basis of TARNet[35]'s causal relationship diagram structure, Do-calculus derivations are performed for y and t in the inference network, respectively, and z and t in the model network to fit the interaction between potential confounding variables and treatment effects. Overall, the following prediction function is involved in the causal variational autoencoder:

$$\mathcal{F}_{\text{CEVAE}} = \mathcal{L} + \sum_{i=1}^{N} \left(\log q\left(t_i = t_i^* \mid \mathbf{x}_i^*\right) + \log q\left(y_i = y_i^* \mid \mathbf{x}_i^*, t_i^*\right)\right) \tag{14}$$

the training set comprises observations of an input variable, treatment variable, and outcome variable at specific data points $\mathbf{x}_i^*$, t_i^*, and y_i^*.

On the ground of CEVAE[45], Sun et al. proposed the LaCIM[95] latent causal model to avoid false associations and improve the generalization ability of the model; CEGAN[61] used GAN networks to identify the potential confounders unbiasedly.

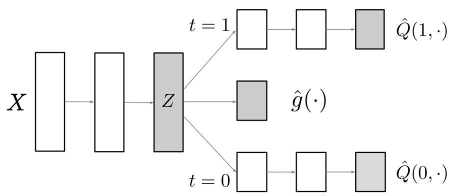

Fig. 5 **Dragonnet**[59]: Propensity score adaptive neural network

Proposed by Shi et al., first-ever Dragonnet[59] adhibited the regularization objective functions into nonparametric estimation theory and the Propensity score prediction networks to CFRNet[35], thus making sure that the covariates are adjusted for treatment-related information in them. As can be seen, Figure 5 shows the network structure of the propensity score adaptive neural network.

In accordance with Dragonnet[59], VSR[96] proposed a reweighting model that removes the association processing and confounding factors. It also used a deep neural network to aggregate the density ratios of latent variables across the variational distribution for calculateing the sample weight distribution; As part of the estimation process, DONUT[99] has added an orthogonal constraint to the non-confounding factors in the total loss; In addition, an end-to-end regularization and reparameterization method called FlexTENet[100] was proposed to learn a new architecture using multi-tasking framework, by which the shared function between causal structures is obtained adaptively. In DIRECT-ND[113], the entangled representation is solved through hybrid learning, and the multivariate causal effect estimation problem was studied from a new perspective. Additionally, VAE and GAN networks were added to the model to realize the learning of hybrid representation space.

Aiming at balancing selection bias, Zhang et al. [87] proposed the RSB algorithm using autoencoder networks via PCC regularization and instrumental variables, and also adding the confounding variables and moderators for prediction. As extention of CFRNet[35], both , DR-CFR[88] and DeR-CFR[58] were proposed with the main structure as shown in Figure 6 to remove the covariate correlations.

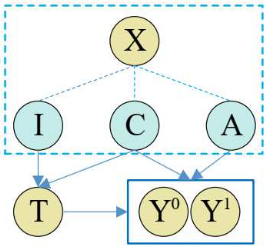

Fig. 6 Decomposed Representation for CounterFactual Regression (DeR-CFR) [58]: Causal framework with decomposed latent factors

As depicted in Figure 6, the observed covariates X can be influenced by three distinct factors: I, an instrumental factor that solely affects the treatment T; C, a confounding factor that impacts both the treatment T and the outcome Y; and A, an adjustment factor that determines the outcome Y. To learn the decomposed representation for counterfactual reasoning, the following steps are taken[58]:

- Representation networks for decomposing $A(X)$, $C(X)$ and $I(X)$ are supplied to facilitate the discovery of latent factors.
- To address confounding variables, three regularization steps are employed: first, A is derived from X with the condition of $A(X) \perp\!\!\!\perp T$ and $A(X)$ being projected onto Y; second, I is derived from X with the constraint that $I(X) \perp\!\!\!\perp Y \mid T$, and $I(X)$ should be predictive of T; finally, $C(X)$ is balanced across the different study groups simultaneously.
- A distinct regression network is employed for each treatment arm to forecast potential outcomes, with input variables $C(X)$ and $A(X)$ and predicted outcomes Y^0 and Y^1.

In the process of decomposition, the orthogonal regularizer is utilized in the following way:

$$\mathcal{L}_O = \bar{W}_I^T \cdot \bar{W}_C + \bar{W}_C^T \cdot \bar{W}_A + \bar{W}_A^T \cdot \bar{W}_I \tag{15}$$

in order to prevent the representation network from rejecting the inputs, the maximum value of one is imposed on $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$. Furthermore, when a hard decomposition is in place, the orthogonal regularizer ensures that each variable in X is associated with only one representation network.

Based on DeR-CFR, CATE's prediction performance has been assessed using CF-CV[114], which selects the best model or hyperparameters from potential candidates. In [115], a meta-learning approach was combined together with deep networks, theoretical reasoning, and the optimal counterfactual information.

Causative treatment effect estimation has always been concerned with rationally using confounding variables. The decoupling of covariates to learn related confounding variables can help remove selection bias and generate unbiased output estimates. Despite its theoretical nature, this kind of methods has also some limitations in practical applications, as it requires decomposing covariates into reasonable explanations.

5.3 GANs-based counterfactual simulation

In deep generative models, generative adversarial networks can capture the uncertainty of counterfactual distributions. The generator produces counterfactual results or consistency distribution between control and treatment groups, while the discriminator fits the unbias estimation of the treatment effects. As well as using factual data, GAN networks also consider the accuracy of counterfactual results when making causal effect estimations. In light of this, generative adversarial models are increasingly used for causal effect estimation.

The first approach suggested by Yoon et al. is for the GANITE[37] network to generate the counterfactual results based on factual data and pass them to the ITE generator. Figure 7 shows the detailed block diagram of GANITE.

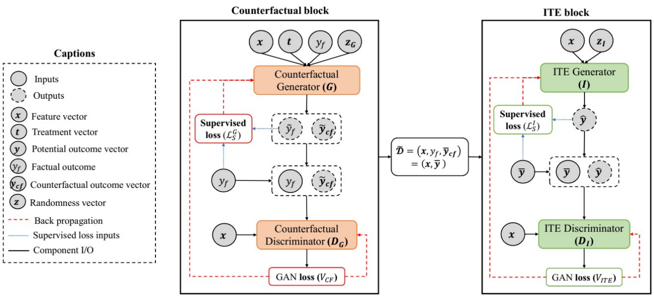

Fig. 7 Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) [37]: Block diagram

For a given feature vector x, GANITE generates the potential output results by first generating factual outputs y_f and then counterfactual samples $\widetilde{y}_{cf}$ using generator G. After these counterfactual data are combined with the original data, a complete data set $\widetilde{D}$ is generated in generator I of the ITE module, which then optimizes $\widetilde{D}$, yielding an unbiased estimate of each treatment's effect.

For the first time, CEGAN[61] has applied the GAN network to balance the distributions between the spatial treatment group and the control group by leveraging the discriminative loss of the GAN network and weighting the Decoder's construct loss or weight after the Encoder. In order to solve the generator-discriminator min-max problem, the following reward function is used:

$$\min_{(\theta_E,\theta_I,\theta_P)} \max_{\theta_D} \mathbb{E}_{q_E(\mathbf{z},\mathbf{x},t,\mathbf{y})}[\log(D(\hat{\mathbf{z}},\mathbf{x},t,\mathbf{y}))] + \mathbb{E}_{q_P(\mathbf{z},\mathbf{x},t,\mathbf{y})}[\log(1-D(\mathbf{z},\mathbf{x},t,\hat{\mathbf{y}}))] \tag{16}$$

these estimates can be expressed as $(D(\hat{\mathbf{z}},\mathbf{x},t,\mathbf{y}))$ and $(1-D(\mathbf{z},\mathbf{x},t,\hat{\mathbf{y}}))$, respectively, as the encoder-decoder's joint distribution, $q_E(\mathbf{z},\mathbf{x},t,\mathbf{y})$ and $q_p(\mathbf{z},\mathbf{x},t,\mathbf{y})$. Different samples belong to different distributions according to the discriminator.

In addition to GANITE and CEGAN, there are also many works using GAN networks to estimate the causal effects in other fields. As part of a generative adversarial framework, GAD[89] applied GAN networks to continuous treatment problems to learn a sample-balanced weight matrix, which removes the association between treatment regimens and covariates; to address multiple treatment as well as consecutive doses treatment problems, DRGAN[90] proposed a model architecture consisting of a couterfactual generator, discriminator, and inference block; As a means of better coping with continuous intervention problems, SCIGAN[62] added a hierarchical discriminator based on DRGAN. In addition, CTAM[26] has also applied the generative adversarial ideas to the treatment effect estimation of text sequence information,. It filters out information related to approximate instrumental variables when learning representations, and matches between the learned representations; To eliminate the association between the treatment and patient history, CRN[91] utilized a counterfactual recurrent neural network to reflect the time-varying treatment effect; In ABCEI[93], the covariate distributions between the control and treatment groups are well balanced with GAN networks, and a regularization function of mutual information estimators is added to reduce bias; To learn the balanced covariate representations, CETransformer[98] combined the attention mechanism with WGAN; In TransTEE[66], Transformer was used for covariate representation, in which the treatment effectiveness is estimated by the Propensity Score Network and the selection bias can be overcome by the GAN network. In particular, the model can also be used for discrete, continuous, structured or dose-related treatments.

In general, it is not hard to extend the ways of individual treatment effect estimation to multiple interventions and continuous dose interventions using the GAN network, and it has a marvelous effect on the balance of representation distribution and the generation of potential results. However, due to lacking a complete theoretical support system, using GAN networks to solve the problem of causal effect estimation requires more impeccable theoretical derivation in the future.

5.4 Time series causal estimation

In treatment effect estimation, most models focus on numerical variables, and it is still difficult to deal with textual information and time series information[116]. Variable decoupling for textual information estimation can reduce estimation bias since there are many covariates in textual information that are unrelated to causal effect estimation. When dealing with time-series information, RNNs have usually been combined to create counterfactual recurrent networks based on historical information.

The R-MSN[64] model is first proposed by Lim et al. in order to address the problems arising with continuous treatment doses and multi-treatments under time series. Figure 8 illustrates the model's frame structure, which uses a recurrent edge network to remove time-dependent confounding, an a standard RNN structure to encode and decode.

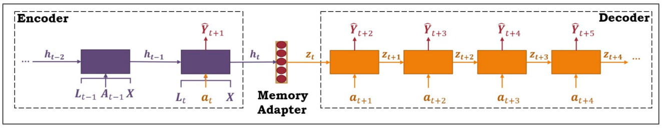

Fig. 8 Recurrent Marginal Structural Network (R-MSN) [64]: Architecture for multi-step treatment response prediction

To predict the causal effect, R-MSN used the standard LSTM structure, dividing multi-treatment and continuous intervention problems according to the corresponding time interval. As a counterfactual recurrent network, CRN[91] constructs a treatment-invariant representation for each time step based on R-MSN[64], eliminating the patient's medical history association between treatment allocation and treatment allocation and balances time-varying confounding biases; By combining with the current treatment assignment and historical information, the hidden confounders are inferred using recurrent weighted neural networks in DSW[110], an then reweighted using time-varying inverse probabilities. In addition to building a multi-task output RNN factor model, TSD[92] allocates multiple treatment over time, an then estimates the treatment effects with multi-cause hidden confounders by which the latent variables free of treatment can be inferred. Furthermore, it substitutes the unobserved confounders with latent variables, and infers logistic regression results in the absence of treatment; In SyncTwin[102], treatment estimation is performed based on the temporal structure of the prediction results, and synthetic twin samples are constructed and counterfactual predictions are obtained.

Yao et al. proposed a matching treatment-adversarial learning CTAM[26] method that takes into account text sequence information. As shown in Figure 9, when learning representations, it filters out the approximate instrumental variables and then makes matching between the learned representations to estimate the treatment effect.

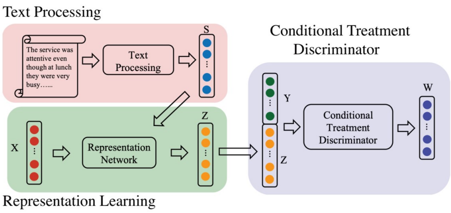

Fig. 9 Conditional Treatment-Adversarial learning based Matching Method (CTAM) [26]: Framework

In particular, CTAM[26] consists of three parts: text processing, representation learning, and conditional treatment discrimination. As the first step, the text processing part converts the original text into a vector representation S, adds non-text covariates X, and then transforms the output into a latent representation Z. As a next step, an arithmetic unit between the representation learning network and the conditioned treatment discriminator is calculated during the training process between Z and Y. In order to filter out instrumental variables, the representation learning network prevents treatment assignment to correlative variables. As a last step, it implements the match in the representation space Z.

To predict treatment assignment using mutual information between global feature representations and individual feature representations, IDRL[117] propsoed to learn Infomax and domain-independent representations. To maximize the ability of capturing the common prediction information among treatment and control groups, the influence of instrumental variables and irrelevant variables were filtered out. In addition, the SCRNet[118] attempted to estimate ITE with different types of variables by dividing the the covariates. Recently, CT[107] adopted transformer and lstm to capture complex long-term dependencies between time-varying confounders and proposed a new counterfactual domain confusion loss to address the confusion bias.

Estimation of causal effects of textual time series is often combined with the problems related to multiprocessing and continuous dose processing. Despite the widespread application of this direction, researchers need to develop a standard for measuring intervention effects based on the actual situation, and it is difficult to assess the rationality and reliability of the various working standards used in the industry.

5.5 Multi-treatment and continuous dose models

Casual estimation for individual treatments focuses on solving binary treatment problems, and extending it to multiple treatment is computationally expensive. However, multiple treatment and continuous dose treatment models have many applications, such as radiotherapy, chemotherapy and surgery for cancer treatment, as well as prolonged usage of vasopressors for many years. It is therefore beneficial to estimating the effects of ongoing interventions in these various treatment settings in order to make marvelous long-term process decisions.

For the first time, Schwab et al. have extended individual treatment estimation to multi-discrete treatment problems with the PM[34] algorithm. In PM , counterfactual reasoning is utilized in small batches by matching nearest neighbors samples, which makes it easily implemented and is compatible with a wide range of architectures, and there is no need to increase computation complexity or other hyperparameters for treating arbitrary quantity of patients. In order to capture higher-order effects, TECE-VAE[80] modeled the dependence between treatments by using task embedding, extending the problem to arbitrary subsets of multi-treatment situations.

When solving problems involving multi-treatments and continuous dose treatments, GAN networks are frequently combined. A two-step generative adversarial de-aliasing algorithm proposed by GAD[89] can be used for continuous treatment problems, removing the association between covariates and treatment variables. Specifically, it is along with the following three steps: A) Creating uncorrelated covariates helps eliminate potential correlations that may introduce bias into the data. B) Addressing sampling bias in the observed data is achieved by assigning sample weights that accurately represent the underlying population. C) Using generative adversarial networks to create synthetic data that is free from biases can improve the quality of the data, leading to more accurate analysis and modeling.

As an improved GAN model, DRGAN[90] takes the form of a generator, a discriminator, and prediction network to generate a complete dose-response curve for each sample by considering both multi-treatment and continuous dose treatment options; By using a hierarchical discriminator on the basis of DRGAN, SCIGAN[62] was proposed to improve the model's ability of handling the problems of continuous intervention.

Meanwhile, a set of open model benchmark parameters, including MISE, DPE, PE, and model selection criteria, were proposed in DRNet[48], which allows the generation of dose-response curves for an unlimited number of treatments under continuous dose parameters. In VCNet[49] that utilizes a variable coefficient neural network, a continuous ADRF[119–121] estimator is automatically calculated for the continuous activation function, which is helpful of preventing the processing information from being lost. Moreover, the existing target regularization method was also extended to obtain a double robust ADRF curve estimator.

As part of DRNet[48], continuous treatments are divided into blocks and trained separately into hidden layers, which are then nested into each other to construct a piecewise fit of individual dose-response curves; A continuous prediction head of weighted treatment is created by VCNet[49] by paying closer attention to treatment continuity, and optimizing the individual prediction head into a mapping function of covariates that change with treatment. The structure comparison of DRNet and VCNet models is shown in Figure 10.

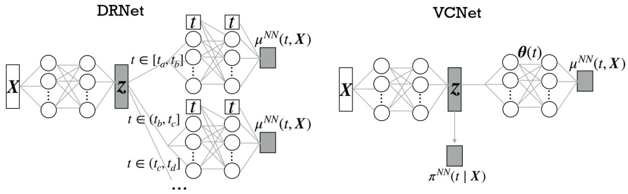

Fig. 10 Dose Response Network (DRNet) [48] and Varying Coefficient Neural Network (VCNet) [49]: Comparison of network structures

Building upon DRnet and VCnet, ADMIT[106] employed a re-weighting scheme and the IPM to estimate an upper bound on the counterfactual loss and obtain the ADRF. Both theoretical and experimental evidence supported the approach. TransTEE[66] combined the hierarchical discriminator of SCIGAN[62] with the variable coefficient structure of VCNet[49] and inducted the Transformer multi-headed attention mechanism generic framework to extend the causal effect estimation problem to discrete, continuous, structured, and dose-related treatments. Specifically, in MetaITE[122], the concept of meta-learning was introduced to the problem of estimating the effects of multiple therapies. A meta-learner was trained from a set of source treatment groups with sufficient samples, and the model was updated through gradient descent with limited samples in the target treatment group.

As the first study of the multi-treatment combination problem, NCoRE[97] has used the cross-treatment interaction to infer the causal generative processes underlying multiple treatment combinations, in which the counterfactual representations learned in a treatment setting are combined.

To estimate the multi-cause perturbation treatment effect, Prichard et.al proposed the idea of SCP[101] for the first time. To overcome the confounding bias in two steps, a single-cause CATE estimator is first applied to augment observed data and estimate potential outcomes; As a next step, the augmented data set is adapted for covariates to obtain multi-factor unbiased estimators. In addition to illustrating the relationship between single-factor and multiple-factor problems, SCP shows the equivalence of conditional expectations of single-factor interventions and multiple-factor interventions, which can be verified by the following equation:

$$\mathbb{E}_{\alpha}\left(Y\left(a_k, \mathbf{a}_{-k}\right) \mid \mathbf{X}\right)=\mathbb{E}_k\left(Y\left(a_k\right) \mid \mathbf{X}, \mathbf{A}_{-k}\left(a_k\right)=\mathbf{a}_{-k}\right) \tag{17}$$

To expand the dataset, the initial step involves adding the outcome variable $Y\left(a_k\right)$ and the observation variable $\mathbf{A}_{-k}\left(a_k\right)$. This allows for a more comprehensive analysis of the data, as both the results of the event and the specific attributes of the event are considered. As such, by training a supervised learning model on the augmented dataset, it is possible to estimate the expected value on the right side of Eq. 17 as well as the multifactorial intervention effect given by the left side in a way that enhances the generalizability of the estimator. In contrast to SCP, OOSR[123] proposed a prediction model using a reweighting way that puts emphasize on the outcome-oriented treatment.

Recently, more and more researchers have taken interest in the problem of multi-treatment and continuous dose therapy, and have also made significant contributions. Nevertheless, there are still many models in this area that need to be developed. Especially, it is still an urgent issue to solve how to formulate a unified causal effect measurement standard.

6 Current Status of Industrial Applications

Deep learning-based causal effect estimation has been widely applied to data-driven decision making across domains, and plays an important role in incentive allocation, debiased recommendation, precision medicine, and identifying and measuring strategic decision-making. In this section, typical applications in related fields are introduced according to their corresponding data characteristics, we will focus on industrial applications using deep causal models, involving marketing, E-commerce, finance, medicine, economics, and education, and cast concerns on their social value and application value.

6.1 Marketing applications

Towards the goal of maximizing the return on investment, it is crucial to decide the assignment of user-specific incentive under limited budget. Uber and Didi Chuxing have incentivized drivers as well as passengers with commission and coupons to boost ride-sharing business. Mobile payments, such as Apple Pay, Alipay, WeChat Pay, have designed various mobile marketing campaigns to increase the usage of their applications. Amazon and JD Digits attempted to give out discounts to enhance attraction to customers. Incentive allocation relies on treatment effect estimation models to estimate users' purchase probability with different incentives. PCAN[124] was proposed to learn an an unbiased model by leveraging a small set of unbiased data. Specifically, a biased network was built to generate unbiased data representation by controlling the distribution difference to a unbiased network. Offline and online experimental results indicated that this model can alleviate the problem of price-bias and lead significant performance improvement for a real-world marketing campaign. Besides, DESCN[103] was developed to integrate the learning process of the treatment propensity, the true responses, and the treatment effect in a joint way simultaneously. It has been validated in voucher distribution business in Lazada, one of the largest E-commerce operators belonging to the Alibaba Group. Related results indicated that this method is advantageous in both performances of ITE estimation and uplift ranking. For multi-treatment scenarios, Amazon presented MEMENTO [125]. Different treatment types all gained matching representation by minimizing the upper bound of the loss, which is the sum of the factual and counterfactual estimation errors. Besides incentive allocation, deep causal models are also applied to other tasks in marketing. For example, a user retention model UR-IPW[126] was proposed by AntGroup to account for impression-revisit effect, where users can have explicit and implicit interactions with the recommender system. The model made full use of both kinds of interactions from the observed data, and the selection bias caused by user's self-selection was eliminated by inverse propensity weighting.

Practical Application: In bilateral business relations, effective marketing strategy should also motivate merchants as well as customers. For example, to encourage mobile payment activities, merchants get shared incentives with customers after them scanning QR-codes and paying with Alipay. However, making independent optimization for both sides can be nonoptimal. Therefore, [127] and [128] took the mutual influence into account, and used graph neural networks to represent merchants and customers jointly by modeling the underlying bipartite influences. Extensive experimental results demonstrate the effectiveness of the proposed approaches.

6.2 E-commerce applications

Recommender systems play an important role in E-commerce. Traditional recommender systems learn user preference by mining correlation in observational data, resulting in biases including selection bias, exposure bias, position bias, conformity bias, etc. To address this issue, Amazon, Netflix, Criteo, Alibaba, AntGroup, JD, Kuaishou, etc. have all begun to utilize causal effect estimation to extract causality. Datasets and recent works on recommendation debiasing can refer to [129] and [130]. In [74], debiased information bottleneck was proposed by Huawei, which is applicable to various types of bias. The model learnt a biased representation in the training phase, which include a biased part and an unbiased part, and only the unbiased part of representation is used in the test phase to reach more accurate recommendations. As mentioned in [131], Wei et al. studied the issue of popularity bias from a cause-effect perspective, where multi-task learning is leveraged to model the contribution of each cause, and counterfactual inference is utilized to eliminate the effect of item popularity during testing. Following that, ESCM^2[132] further exploit the sequential pattern of user actions to address the data sparsity issue and employed a counterfactual risk minimizer to simultaneously address the issues of estimation bias in CVR estimation and potential independence priority in CTCVR estimation. Furthermore, in some scenarios, another set of uniform data can also be used to improve model performances by reducing the data bias. In [12], Criteo proposed a multi-task objective that jointly factorizes the matrix of biased data and the matrix of uniform data. In [133], a general knowledge distillation framework that enables uniform data modeling was proposed, which performs distillation from the view of label, feature, sample, and model structure. Differently, in [134], AutoDebias employed the technique of bi-level optimization to optimize the debiasing parameters with a small set of uniform data.

Practical Application: In [132], CausalMTA was proposed by Alibaba, which can remove the confounding bias for both static and dynamic perspectives.

It is achieved by sample reweighting, which enables learning an unbiased conversion prediction model. Its effectiveness is demonstrated by a real advertisement impression data.

6.3 Financial applications

In the last few decades, economists have learned to take very seriously the old admonition from undergraduate econometrics that "correlation is not causality". We have surveyed a number of recent developments in the econometrics toolkit for addressing causality issues in the context of estimating the impact of policies. Hennessy et al. [135] analyzed the meaning and utilization of alternative causal effect measures in the types of settings commonplace in finance and economics, those where individual agents have private information and outcome variables are mediated by the beliefs of other agents in the economy. They suggested the utilization of two distinct causal effect definitions which include partial causal effects and total causal effects.

Tiffin [136] focused mostly on the Causal Forest algorithm, and took the costs of a financial crisis as an illustrative example, the paper has endeavored to show how such techniques can produce plausible results, i.e., the estimated average impact of a crisis. A natural application of personalized treatment effect estimation is to estimate optimal policy functions in observational data. Recently, Athey and Wager [137] brought in insights from semiparametric efficiency theory in econometrics to propose a new estimator for optimal policies and to analyze the properties of this estimator. Policies can be compared in terms of their "risk", which is defined as the gap between the expected outcomes using the (unknown) optimal policy and the estimated policy. Arpino et al. [138] proposed an explicitly model, which was applied to the evaluation of a policy implemented in Tuscany on small handicraft firms, to address the violates the Stable Cell Handling Value Assumption due to the interference between cells. Results show that the benefits from the policy are reduced when treated firms are subject to high levels of interference.

Practical Application: In the field of Fin Tech, MYbank of Ant Group adopted causal counterfactual inference debiasing methods [96, 97, 101, 139–141] to solve the biased problem of loan marketing AB experiment. In order to effectively measure the intervention effect, we use the method of causal counterfactual inference to construct a homogeneous control group from the full experimental population based on observational data, so that horizontal comparisons can be made.

6.4 Medical applications

In the area of precision medicine, the research community has started to explore quantitative individual-level effect of a treatment, by using observational data from electronic health records. Clinical decision-makers target at determining the personalized optimal treatment process, using static or time-series observational data. However, observation datasets usually face the problem of treatment assignment bias. To address it, a series of models have been proposed. In many medical applications, multiple variables can be intervened at the same time. In order to estimate impact of multi-cause treatment, Single-cause Perturbation[101] was proposed, which firstly augments the observational dataset with potential outcomes estimated from single-cause interventions, and then adjusts the covariate of the augmented dataset to learn the estimator. Following that, GraphITE[142] was proposed, which utilizes graph neural networks to learn the representation of the graph-structured treatments. In order to reduce the observation biases, HSIC regularization is employed to obtain an independent representation of the targets in regard to the treatments. Observational data also contain information on complex time-varying treatment. To estimate the effectiveness of treatment over time, SCIGAN[62] was proposed, which is flexible and has the ablity to estimate counterfactual outcomes for sequential interventions. Additionally, TE-CDE[143] was proposed, which can estimate the potential outcomes at any time point. To do so, it uses adversarial training to adjust for time-dependent confounding.

Practical Application: The van der Schaar Lab has a wide range of clinical applications on caulity models, including COVID-19, organ transplantation, etc. Particularly, to decide "one best model" for each scenario, a validation procedure is introduced in [144] to estimate the performances of different causal effect estimation methods by using influence functions.

6.5 Economic applications

Causal effect estimation has attracted more and more attention of researchers in economic science, especially how to better combine with deep learning to solve practical problems. Learning about cause and effect is arguably the main goal in applied econometrics.

For instance, unobserved confounding factors threaten the internal validity of estimates, data availability is often limited to non-random, selection-biased samples, causal effects need to be learned from surrogate experiments with imperfect compliance, and causal knowledge has to be extrapolated across structurally heterogeneous populations. A powerful causal effect estimation framework is required in order to tackle all of these challenges, which plague essentially any data analysis to varying degrees [145]. It nests the prediction power of deep learning to obtain consistent causal estimators under high dimensional covariates [146–148]. In [149], Angrist and Steve Pischke coined the term "credibility revolution". They argued that economics turns to transparent empirical strategies applied to specific causal problems. Problem-driven methodological agendas are mostly based on the propensity score theorem of Rosenbaum and Rubin [150]. This theorem changes applied econometrics to focus our attention on the process of determining treatment assignment rather than models for outcomes. Dehejia and Wahba [77] was the first to demonstrate the value of this approach. More recently, Belloni et al. [151] utilized deep learning to model scores while modeling outcomes. This work can be seen as an extension of Robins's notion of dual robustness to a broader class of empirical strategies. Angrist [152] introduced a novel framework, the local average treatment effects framework for causal effect estimation, to help make the empirical strategies in economics should be transparent and credible. The LATE theorem can tell us for whom particular instrumental variables and regression discontinuity estimates are valid. Hal R. Varian [153] argued that the powerful techniques used in deep learning may be useful for developing better estimates of the counterfactual, potentially improving causal effect estimation. For generalized neighbor matching to estimate individual and average treatment effects, Vikas [154] proposed the use of deep learning techniques in econometrics, specifically for causal effect estimation and for estimating individual as well as average treatment effects.

6.6 Educational applications

Education plays a central role in modern society, especially in earlier years [155], a wave of new studies on the effects of educational interventions on student performance has emerged. Therefore, the need for evidence-based policy in the field of education is increasingly recognized. Education policymakers and practitioners want to know which policies and practices can best achieve their goals. However, providing empirical evidence suitable for guiding policy is not an easy task, because it refers to causal effect estimations that require special research methods which are not always easy to communicate due to their technical complexity. From a Bayesian perspective, David [156, 157] introduced a review and synthesis of the problem of causal inference in large scale educational assessments which requires the articulation of framework for causal effect estimation followed by a statistical approach that closely matches the framework and can yield the causal estimate of interest. Zhao et al. [32] proposed the Residual Counterfactual Networks in an Intelligent Tutor System can decide which hint is more suitable for a specific student. However, the effectiveness of an intervention is necessarily multifaceted and complex effects differ between students, as a function of implementation [158], and, potentially, as a function of time and location. Sales et al. [159] explored a different sort of treatment effect heterogeneity differences in effectiveness for different outcomes. Specifically, different posttest items measuring different skills. Carvalho [160] conductd causal effect estimation on students' online behavior patterns using a different toolkit, GeNIe3, using data from a learning management system. Chen et al. [161] developed a causal discovery framework that utilized TETRAD [162].

7 Guideline for Experiment

After an in-depth description of the deep causal modeling approach, this section will provide a detailed experimental guide, including a comprehensive conclusion and analysis of datasets, source codes as well as experiments.

7.1 Datasets

As counterfactual results can never be observed in real life, finding datasets that satisfy experimental requirements is difficult. Most of the datasets used in the literature are semi-synthetic. Following the survey[2], we further supplemented in Table 3 more relevant datasets that are popularly used for causal analysis. Meanwhile, we also conclude the application scenarios for these datasets. In addition, Table 4 summarizes the web links to these datasets and the classical models that are evaluated on them. Below are the detailed descriptions of the available datasets.

IHDP. The Infant Health and Development[78] dataset originated from a controlled trial that randomly assigned preterm infants with low birth weight to either a treatment or a control group. The study collected pre-treatment covariates such as birth weight, head circumference, neonatal health index, and maternal characteristics like age, education, drug, and alcohol use. The treatment group received intensive and high-quality childcare, including specialized home visits[163], while the control group received standard care. The primary outcome measure was the cognitive test score of the infants. To develop unbiased selection models, the study excluded noisy subsets of the treatment group from the analysis.

Jobs. In the Jobs in LaLonde (1986)[79] study, the employment dataset was utilized to assess the effect of vocational training on employment outcomes. The dataset consisted of both randomized data from state-supported work programs and non-randomized data from observational studies. Pre-treatment covariates, such as age, education, race, and income, were collected for the years 1974 and 1975. The treatment group received vocational training, while the control group did not receive any training. The primary focus was on the employment status of individuals in both groups as the primary outcome variable.

Twins. The Twins dataset consists of data on twin births in the United States between 1989-1991[164], including 40 pre-treatment covariates related to pregnancy, twin births, and parents. These covariates include gestational weeks, quality of care during pregnancy, pregnancy risk factors such as anemia, alcohol, and smoking, nursing, and residence. The primary outcome measure for each twin pair is the one-year mortality rate. The dataset also features a twin dataset, which categorizes twins into treatment (heavier twin) and control (lighter twin) groups. To simulate selection bias, the study assigns different treatments based on user-defined criteria.

News. The News dataset is a sample of 5000 news articles randomly selected from the New York Times corpus. The dataset explores media consumers' perceptions of these news items and includes word counts for each article. The primary outcome measure is readers' opinions of the news item, and the treatments investigated are different devices used to view the news, such as smartphones, tablets, computers, and TVs. The dataset offers valuable insights into how consumers interact with news articles on various devices and can inform media companies' content delivery strategies.

ACIC. Every year since 2016, the Atlantic Causal Inference Conference has organized a challenge focused on estimating causal effects from various datasets. The latest conference dataset for this challenge is summarized in a publication titled "ACIC"[165].

TCGA. The Cancer Genome Atlas (TCGA)[83] is a comprehensive and extensive database of genomic information, containing billions of genetic sequences. It includes data from 9658 individuals who have undergone different cancer treatments, such as drug therapy, chemotherapy, and surgical procedures, with the aim of assessing the risk of cancer recurrence or development after treatment.

PK-PD model of tumor growth. The pharmacokinetic-pharmacodynamics (PK-PD)[166] model is a valuable tool for exploring dose-response relationships and guiding treatment decisions. This model can be particularly useful in the treatment of non-small cell lung cancer, where factors such as the combined effects of chemotherapy and radiotherapy, cellular regeneration post-treatment, patient prognosis, and variations in tumor size at diagnosis must be considered. By analyzing expected treatment responses[167, 168], the PK-PD model can predict tumor growth in the presence of time-dependent confounding variables. This model is widely used by clinicians to generate hypotheses and identify optimal treatment options.

MIMIC III. MIMIC III[169] is an electronic health records database that contains data from ICU patients. The benchmark dataset comprises 7413 samples and 25 covariates after filtering out missing values. In the ICU, common treatment options for sepsis patients include antibiotics, vasopressors, and mechanical ventilators. Vital signs such as white blood cell count, blood pressure, and oxygen saturation are used over time to evaluate the effects of these treatments on patients. For a detailed description of the clinical data, please refer to the paper[170].

NICO. The NICO image dataset used for object classificationn[95] may exhibit bias in sample selection. The "animals" dataset, which is often used for cat or dog classification, is considered a benchmark distribution for non-i.i.d. data. However, this bias may arise due to various factors such as the timing and criteria for selecting images, the context in which animals are presented, and the semantic shape of the animals. Additionally, environmental factors like the presence of "grass" or "snow" can also influence classification results.

ADNI. The ADNI[171] dataset contains three latent representation outputs: Alzheimer's Disease, Mild Cognitive Impairment, and Normal Control. Covariates in the dataset include age and TAU[172], which can be utilized to determine the suitability of Magnetic Resonance Imaging (MRI) as a therapeutic input.

CHESS. The CHESS dataset was designed to collect information about ICU patients during the first wave of the COVID-19[81, 82] pandemic in England. The dataset includes risk factors, treatments, and outcomes for 3090 patients. Some of the covariates in the dataset are age, multiple morbidity, ventilation, and antiviral drug treatments. The primary outcome measured in the dataset is the length of stay in the ICU[173].

Table 3 Overview of some open available datasets for causal analysis and their application scenario

Datasets	Binary	Multiple	Continuous dose	Times series	Structured
IHDP	✓	✓			
Jobs	✓				
Twins	✓				
News	✓	✓	✓		
TCGA			✓		✓
ACIC		✓	✓		
Tumor Growth				✓	
MIMIC III			✓		
NICO	✓				
CMNIST	✓				
ADNI		✓			
CHESS		✓			
CPRD				✓	
BlogCatalog		✓			✓
Flickr		✓			✓

CPRD. The Clinical Practice Research Datalink (CPRD)[174] is a database containing medical records from NHS general practice clinics in the UK, covering around 6.9 % of the country's population. It has been found to be associated with secondary care admissions based on national mortality records and hospital event statistics. The study involves measuring low-density lipoprotein levels after CPRD initiation, with treatment defined as the date of the first prescription. Prior to treatment initiation, the following risk factors are measured as temporal covariates: high-density lipoprotein cholesterol, blood pressure, pulse, creatinine, triglycerides, and smoking status. Participants for the HPS registry are selected from a pool of 125,784 eligible individuals. The treatment and control groups are divided into three equally sized subsets for training, validation, and testing, resulting in a total of 17,371 treatment groups and 24,557 control groups.

BlogCatalog. BlogCatalog[175] is an online community where bloggers can post their content, and social relationships between bloggers are represented as edges in the dataset. Each blogger is represented as an instance, and the blog descriptions are represented using keywords in a bag-of-words format. By analyzing reader opinions, the study investigates whether blog content gets more comments on mobile or desktop devices. The study uses the content of readers' comments on mobile devices versus desktop devices to estimate parameters for individual treatment effects. If a blogger's content is read more on a mobile device than on a desktop device, the blogger belongs to the treatment group, and vice versa.

Flickr. Flickr[176] is an online social networking platform where users can upload and share photos and videos. The dataset comprises instances representing individual users and edges that indicate their social relationships. Each user's features are represented by tags of interest. The study's settings and assumptions are similar to those of the BlogCatalog dataset, with the aim of identifying which tags are associated with increased engagement on the platform. The study also seeks to estimate individual treatment effects using parameters derived from user comments and engagement with photos and videos.

Table 4 The web links to the open available datasets and the related models that have used them for performance evaluation

Datasets	Links	Methods
IHDP	https://www.fredjo.com	[16, 26, 33–37, 45–47, 49, 58, 59, 66, 84, 85, 87, 88, 93, 94, 98, 98, 100, 115, 118]
Jobs	https://www.fredjo.com	[35–37, 45–47, 65, 93, 98, 99, 104, 105, 109, 117, 118]
Twins	www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html	[36, 37, 45–47, 61, 89, 93, 98, 99, 104, 105, 109]
News	https://archive.ics.uci.edu/ml/datasets/bag+of+words	[26, 33, 34, 48, 49, 62, 65, 66, 90, 106, 117, 122]
ACIC	https://jenniferhill7.wixsite.com/acic-2016/competition	[59, 93, 100, 103, 177–180]
TCGA	https://gdc.cancer.gov/	[62, 65, 66, 90, 106, 123, 181–185]
Tumor Growth	www.nature.com/scientificreports/	[64, 91, 186–189]
MIMIC III	https://mimic.physionet.org/	[48, 62, 65, 92, 93, 107, 190]
NICO	https://www.dropbox.com/sh/8mouawi5guaupyb/AAD4fdySrA6fn3PgSmhKwFgvadl=0	[95, 191–195]
ADNI	https://adni.loni.usc.edu/	[95, 196–199]
CHESS	https://www.heywhale.com/mw/dataset/5e8ee81fe7ec38002d00f9cb	[101, 200–202]
CPRD	https://academic.oup.com/ije/article/44/3/827/632531	[102, 203–207]
BlogCatalog	https://www.blogcatalog.com	[111, 208, 209]
Flickr	https://www.flickr.com	[111]

7.2 Source codes

Most existing reviews are limited to an overview of traditional algorithm code. Here we summarize the source code associated with deep learning and programming frameworks used. In academia and industry, PyTorch and TensorFlow are the mainstream deep learning frameworks, with more than 80 % scholars and programmers using them for model construction. In addition to the popularly used datasets as motioned above, we also list in Table 5 the available source codes for some representative models and the relevant dataset they have used.

Table 5 Overview of some available datasets and web links to them

Methods	Datasets	Framework	Links
DCN-PD[84]	IHDP	Pytorch	https://github.com/Shantanu48114860/Deep-Counterfactual-Networks-with-Propensity-Dropout
BNN[33],CFRNet[35]	IHDP,Jobs,News	Tensorflow	https://github.com/clinicalml/cfrnet
CEVAE[45]	IHDP,Twins,Jobs	Tensorflow	https://github.com/AMLab-Amsterdam/CEVAE
GANITE[37]	IHDP,Twins,Jobs	Tensorflow	https://github.com/jsyoon0823/GANITE
SITE[36]	IHDP,Twins,Jobs	Tensorflow	https://github.com/Osier-Yi/SITE
R-MSN[64]	PK-PD model of tumor growth	Tensorflow	https://github.com/sjblim/rmsn_nips_2018
PM[34]	IHDP,News	Tensorflow	https://github.com/d916b/perfect_match
Dragonnet[59]	IHDP,ACIC	Tensorflow	https://github.com/claudiashi57/dragonnet
DKLITE[47]	IHDP,Twins,Jobs	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/dklite
CRN[91]	PK-PD model of tumor growth	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/counterfactual_recurrent_network
TSD[92]	MIMIC III	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/time_series_deconfounder
ABCEI[93]	IHDP,Twins,Jobs,ACIC,MIMIC III	Tensorflow	https://github.com/octeufer/Adversarial-Balancing-based-representation-learning-for-Causal-Effect-Inference
LaCIM[95]	NICO,CMNIST,ADNI	Pytorch	https://github.com/wubotong/LaCIM
SCIGAN[62]	TCGA,News,MIMIC III	Tensorflow	https://github.com/ioanabica/SCIGAN
DRNet[48]	TCGA,News,MIMIC III	Tensorflow	https://github.com/d909b/drnet
VCNet[49]	IHDP,News	Pytorch	https://github.com/lushleaf/varying-coefficient-net-with-functional-tr
DeR-CFR[210]	IHDP,Jobs,Twins	Tensorflow	https://github.com/anpwu/DeR-CFR
DONUT[99]	IHDP,Twins,Jobs	Tensorflow	https://github.com/tobhatt/donut
FlexTENet[100],CATENets[115]	IHDP,Twins,ACIC	Jax,Pytorch	https://github.com/AliciaCurth/CATENets
SCP[101]	COVID-19	Pytorch	https://github.com/vanderschaarlab/Single-Cause-Perturbation-NeurIPS-2021
SyncTwin[102]	CPRD	Pytorch	https://github.com/vanderschaarlab/SyncTwin-NeurIPS-2021
TransTEE[66]	IHDP,News,TCGA	Pytorch	https://github.com/hlzhang109/TransTEE
CF-CV[114]	IHDP	Tensorflow	https://github.com/usaito/counterfactual-cv
CGN[63]	MNIST,ImageNet	Pytorch	https://github.com/autonomousvision/counterfactual_generative_networks
DESCN[103]	ACIC,Epilspsy	Pytorch	https://github.com/kailiang-zhong/DESCN
CT[107]	MIMIC III	Pytorch	https://github.com/Valentyn1997/CausalTransformer
CBRE[104]	IHDP,Twins,Jobs	Tensorflow	https://github.com/jameszhou-gl/CBRE
CITE[105]	IHDP,Twins,Jobs	Tensorflow	https://github.com/XinshuLI2022/CITE
ADMIT[106]	TCGA,News	Pytorch	https://github.com/waxin/ADMIT

By combining related methods, datasets, and source codes, we can more easily identify the innovation points in each model. Meanwhile, it will also facilitate more fair comparison in performance evaluation. In addition, undoubtedly, these source codes will also greatly promote the development of research community about causal effect estimation. As an example, the covariate decomposition is applied to the Dragonnet[59] model when combined with the DeR-CFR[58] model to make a further model optimization. By applying the TransTEE[66] attention mechanism to the representation balance part of VCNet[49] or DRNet[48], the continuous dose estimation curve can be fitted more accurately. It also means the latest advances in causal analysis have also benefited from or inspired by some previous representative works.

7.3 Experiment

We also provide an experimental summary of the binary treatment, multiple treatment and continuous dose treatment scenarios on the above datasets, respectively. it should be pointed out that the reported results are under unified dataset setting for the compared models. Detailed experimental results and performance comparisons can be found in Tables 6, 7, 8, and 9, respectively, in which the **Mean** $\pm$ **Std** of the estimated treatment effect results (lower better) are presented.

7.3.1 Settings and results for binary treatment

The study utilized three different binary treatment datasets, namely IHDP[78], Jobs[79], and Twins[164].To obtain the results, the IHDP and Twins datasets were subjected to 10 realizations, with a 63/27/10 split ratio for train/-validation/test sets. On the other hand, the Jobs dataset underwent 10 train/validation/test splits with split ratios of 56/24/20. The study presented the experimental results in Tables 6, 7 and 8, showing the average performance across the different splits and realizations.

IHDP: To create an unbalanced treatment group, a subset of the population was intentionally removed before the experiment, resulting in an uneven distribution. The simulated results were obtained using the "A" setting of the NPCI package, and the real effect was calculated using noise-free results, similar to the methodology used in[163]. To further explore the impact of increasing imbalance between the original treatment groups, a biased subsample of the IHDP dataset was constructed. This allowed for a more nuanced analysis of the effects of growing imbalance, adding to the understanding of the overall impact on the results of the experiment.

Twins: The study's dataset focused on same-sex twin couples weighing less than 2,000 grams, resulting in 5,409 records after removing those with missing features. To create a selection bias, the study selected one twin to observe and hid the other. This was done using the procedure $T_i \mid \mathbf{x}_i \sim \text{Bern}\left(\text{Sigmoid}\left(\mathbf{w}^T\mathbf{x} + n\right)\right)$, where $\mathbf{w}^T$ was randomly chosen from a uniform distribution $(-0.1, 0.1)$ with 40×1 size, and $\mathcal{N}$ was added noise from a normal distribution with mean 0 and standard deviation 0.1.

Jobs: The Jobs dataset is designed for binary classification, with the aim of predicting unemployment based on a set of features described in[211]. The dataset includes a LaLonde trial sample[212] with 297 treatment and 425 control observations, as well as a PSID comparison group with 2490 control observations. At the end of the study, the dataset contained 482 unemployed subjects, accounting for 15% of the total population.

7.3.2 Settings and results for multiple and continue dose treatments

In the field of multiple and continuous dose treatments, researchers have relied on three publicly available datasets: TCGA[83], News[33] and MIMIC

Table 6 Overview of performance comparisons on IHDP dataset by some representative models in the case of binary treatment

Methods	IHDP(ϵ_{ATE})		IHDP($\sqrt{\epsilon_{PEHE}}$)	
	In-sample	Out-sample	In-sample	Out-sample
BNN[33, 35]	0.37 ± 0.03	0.42 ± 0.03	2.2 ± 0.1	2.1 ± 0.1
TARNet[35]	0.26 ± 0.01	0.28 ± 0.01	0.88 ± 0.02	0.95 ± 0.02
CFR_{MMD}[35]	0.30 ± 0.01	0.31 ± 0.01	0.73 ± 0.01	0.78 ± 0.02
CFR_{WASS}[35]	0.25 ± 0.01	0.27 ± 0.01	0.71 ± 0.02	0.76 ± 0.02
CEVAE[45]	0.34 ± 0.01	0.46 ± 0.02	2.7 ± 0.1	2.6 ± 0.1
GANITE[37]	0.43 ± 0.05	0.49 ± 0.05	1.9 ± 0.4	2.4 ± 0.4
SITE[36]	———	———	0.604 ± 0.093	0.656 ± 0.108
ACE[46]	———	———	0.489 ± 0.046	0.541 ± 0.061
DKLITE[47]	———	———	0.52 ± 0.02	0.65 ± 0.03
DR-CFR[88]	0.240 ± 0.032	0.261 ± 0.036	0.657 ± 0.028	0.789 ± 0.091
CETransformer[98]	———	———	0.46 ± 0.02	0.51 ± 0.03
DeR-CFR[58]	0.130 ± 0.020	0.147 ± 0.022	0.444 ± 0.020	0.529 ± 0.068
DONUT[99]	0.13 ± 0.01	0.19 ± 0.02	———	———
NESTER[109]	0.06 ± 0.04	0.09 ± 0.07	0.73 ± 0.19	0.76 ± 0.20
CBRE[104]	0.10 ± 0.01	0.13 ± 0.02	0.52 ± 0.00	0.60 ± 0.1
CITE[105]	0.09 ± 0.01	0.11 ± 0.02	0.58 ± 0.1	0.60 ± 0.1

Table 7 Overview of performance comparisons on Twins dataset by some representative models in the case of binary treatment

Methods	Twins($\hat{\epsilon}_{ATE}$)		Twins($\sqrt{\hat{\varepsilon}_{PEHE}}$)	
	In-sample	Out-sample	In-sample	Out-sample
BNN[33, 35]	0.0056 ± 0.0032	0.0203 ± 0.0071	0.307 ± 0.001	0.309 ± 0.004
TARNet[35]	0.0108 ± 0.0017	0.0151 ± 0.0018	0.314 ± 0.001	0.313 ± 0.002
CFR_{MMD}[35]	———	———	0.312 ± 0.001	0.316 ± 0.003
CFR_{WASS}[35]	0.0112 ± 0.0016	0.0284 ± 0.0032	0.308 ± 0.001	0.309 ± 0.003
GANITE[37]	0.0058 ± 0.0017	0.0089 ± 0.0075	0.289 ± 0.005	0.297 ± 0.016
SITE[36]	———	———	0.309 ± 0.002	0.311 ± 0.004
ACE[46]	———	———	0.306 ± 0.000	0.301 ± 0.002
DKLITE[47]	———	———	0.288 ± 0.001	0.293 ± 0.003
CETransformer[98]	———	———	0.287 ± 0.001	0.289 ± 0.002
DONUT[99]	0.0025 ± 0.0016	0.0033 ± 0.0026	———	———
NESTER[109]	0.0034 ± 0.0026	0.063 ± 0.0033	0.318 ± 0.002	0.319 ± 0.000

Table 8 Overview of performance comparisons on Jobs dataset by some representative models in the case of binary treatment

Methods	Jobs(ϵ_{ATT})		Jobs($\mathcal{R}_{pol}(\pi_f)$)	
	In-sample	Out-sample	In-sample	Out-sample
BNN[33, 35]	0.04 ± 0.01	0.09 ± 0.04	0.20 ± 0.01	0.24 ± 0.02
TARNet[35]	0.05 ± 0.02	0.11 ± 0.04	0.17 ± 0.01	0.21 ± 0.01
CFR_{MMD}[35]	0.04 ± 0.01	0.08 ± 0.03	0.18 ± 0.00	0.21 ± 0.01
CFR_{WASS}[35]	0.04 ± 0.01	0.09 ± 0.03	0.17 ± 0.01	0.21 ± 0.01
CEVAE[45]	0.02 ± 0.01	0.03 ± 0.01	0.15 ± 0.00	0.26 ± 0.00
GANITE[37]	0.01 ± 0.01	0.06 ± 0.03	0.13 ± 0.01	0.14 ± 0.01
SITE[36]	———	———	0.224 ± 0.004	0.219 ± 0.009
ACE[46]	———	———	0.216 ± 0.005	0.215 ± 0.009
DKLITE[47]	———	———	0.13 ± 0.01	0.14 ± 0.01
CETransformer[98]	———	———	0.12 ± 0.01	0.13 ± 0.00
DONUT[99]	0.01 ± 0.00	0.06 ± 0.05	———	———
SCI[65]	———	———	0.204 ± 0.008	0.225 ± 0.014
DeR-CFR[210]	0.053 ± 0.084	0.093 ± 0.032	0.187 ± 0.037	0.208 ± 0.062
NESTER[109]	0.06 ± 0.00	0.02 ± 0.01	———	———
CBRE[104]	0.10 ± 0.03	0.21 ± 0.07	0.13 ± 0.00	0.28 ± 0.00
CITE[105]	0.06 ± 0.02	0.07 ± 0.03	0.23 ± 0.02	0.88 ± 0.0

III[169]. These datasets were partitioned into 64/16/20% for training, validation, and testing. Results obtained from experiments conducted using TARNet, TransTEE, and VCNet on the TCGA dataset, as shown in Table 9, were obtained through the re-implementation of the source code.

The experiment[62, 90] involved administering three treatments, each with a unique dose and associated with a set of parameters. These parameters were randomly sampled from a normal distribution and normalized. Gaussian noise was added to the results obtained from the experiment. The dose for each treatment was assigned by drawing from a beta distribution, and the treatment assignment was based on a categorical distribution with a softmax function of a κ value. The degree of selection bias could be adjusted by changing the κ value.

TCGA: The experiment prioritized the 4000 most variable genes and scaled each gene expression feature within the $[0, 1]$ range. To make the treatment and dose more relevant, the experiment categorized them as either chemotherapy, radiotherapy, or immunotherapy. This approach aimed to provide a more interpretable and clinically relevant experimental design[48].

News: During the experiments, researchers gathered a dataset consisting of 10,000 news articles. Each article was composed of 2,858 features. The two variables, "treatment" and "dose," were used to examine the impact of different viewing devices and reading times on the articles. Treatment referred to the type of device used to view the article, such as a phone or tablet. Dose referred to the amount of time spent reading the article[35, 48].

MIMIC: In this study, 3000 patients who received antibiotics treatment were analyzed. The researchers collected 9 clinical covariates, including age, temperature, heart rate, systolic and diastolic blood pressure, SpO2, FiO2, glucose, and white blood cell count, which were measured at the beginning of the patients' ICU stay. All features were scaled to fit within the range of 0 to 1. The study investigated the impact of different antibiotics and their doses, which were considered as treatment variables[62].

Table 9 Overview of performance comparisons on TCGA, News, and MIMIC datesets by some representative models in the case of multiple and continuous dose treatments

Methods	TCGA			News			MIMIC		
	MISE	DPE	PE	MISE	DPE	PE	MISE	DPE	PE
TARNet[35]	5.76 ± 0.15	0.53 ± 0.06	0.65 ± 0.07	—	—	—	—	—	—
DRN-W[48]	3.71 ± 0.12	0.50 ± 0.05	0.63 ± 0.05	5.07 ± 0.12	4.21 ± 0.11	4.56 ± 0.12	4.47 ± 0.12	0.53 ± 0.05	1.37 ± 0.05
DRNet[48]	3.64 ± 0.12	0.51 ± 0.05	0.67 ± 0.05	4.98 ± 0.12	4.39 ± 0.11	4.17 ± 0.11	4.45 ± 0.12	0.52 ± 0.05	1.44 ± 0.05
DRGAN,SCIGAN[62]	1.89 ± 0.05	0.31 ± 0.05	0.25 ± 0.05	3.71 ± 0.05	4.14 ± 0.11	3.90 ± 0.05	2.09 ± 0.12	0.51 ± 0.05	0.32 ± 0.05
VCNet[49]	6.36 ± 0.11	0.22 ± 0.04	0.32 ± 0.02	—	—	—	—	—	—
TransTEE[66]	6.40 ± 0.14	0.17 ± 0.03	0.78 ± 0.05	—	—	—	—	—	—
TransTEE+TR[66]	6.26 ± 0.13	0.08 ± 0.02	0.96 ± 0.06	—	—	—	—	—	—
TransTEE+PTR[66]	6.36 ± 0.14	0.18 ± 0.03	0.81 ± 0.05	—	—	—	—	—	—

Based on the experimental analysis, we found that with the development and evolution of the deep model, it has significant implications on solving the core issues and challenges such as representation learning, de-biasing, and counterfactual inference for causal effect estimation. Whether it is in the decision making of therapeutic interventions, the fitting of dose curves or the capture of time-varying confounding, deep causal models give us a new perspective to explore a broader range of directions.

8 Conclusions and Future Prospect in Industrial Application

8.1 Conclusions

Deep causal models have become increasingly popular as a research topic because of the development of causal effect estimation and deep learning. It is possible to improve causal effect estimation accuracy and unbiasedness by applying deep network models to causal effect estimation. Moreover, the deep network can be optimized and improved by the profound theory used in causal effect estimation. This survey presents the development of deep causal models and the evolution of various methods. Firstly, the basic knowledge related to the field of causal effect estimation is adhibited. Then, we present the classical treatments and metrics. Additionally, we provide a comprehensive analysis of the deep causal model from temporal development. Next, we divide the deep causal modeling methods into five groups with an overview and analysis. Furthermore, We furnish a comprehensive conclusion of the application of causal effect estimation in industry. Finally, we summarize the relevant benchmark datasets, open source codes, and performance results as experimental guidelines.

Since 2016, causal effect estimation has been combined with deep learning models for the first time in the binary treatment case for estimation of counterfactual outcomes. So far, deep causal models have been used for time-series, multivariate treatment, and continuous dose treatment situations. It is inseparable from the proposal of deep network models such as AE, GAN, RNN, and Transformer by researchers in the field of deep learning, the generation and simulation of datasets such as IHDP, Twins, Jobs, News, and TCGA by researchers in the field of statistics, and the exploration of ATE, PEHE, MISE, DPE by researchers in industry guided with the theory of Potential Outcome Framework. We believe that with the joint efforts of everyone in the community of causal learning, deep causal models will flourish for the benefit of society and humanity.

8.2 Future prospect in industrial application

For marketing applications, in addition to incentive efficacy evaluation scores, the interpretability of deep learning is also necessary to understand why we predict the output and use it as a basis for business innovation. Furthermore, marketing decisions may involve ethical and legal issues, such as loan applications. Therefore, it makes sense to use causal effect estimation as a guarantee of fairness.

For e-commerce applications, existing debias methods in recommendation systems are usually designed to address only one or two specific biases. There is an urgent need for a universal debias framework to deal with all kinds of prejudice. In addition, how to evaluate a recommendation system fairly and impartially is also an important issue. Existing methods either require accurate propensity scores or rely on unbiased data. Therefore, deep causal models are urgently needed to provide theoretical assurance.

For financial and economic applications, it is desirable to integrate both micro- and macro-level data to study causal effects for stability. In addition, uncertainty quantification is also beneficial to decision making processes, which can be realized by Bayesian approximation and ensemble learning techniques with deep learning.

For medical and educational applications, deep causal models can be adopted in a variety of fields including processing high-dimensional data, enriching randomized trials with real-world data, evaluating spillover causal effects, and moving from studies conducted on specific populations to other populations of interest.

References

[1] Guo, R., Cheng, L., Li: A survey of learning causality with data: Problems and methods. ACM Computing Surveys (CSUR) **53**(4), 1–37 (2020)

[2] Yao, L., Chu, Z., Li: A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD) **15**(5), 1–46 (2021)

[3] Li, S., Vlassis, N., Kawale: Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In: IJCAI, pp. 3768–3774 (2016)

[4] Fong, C., Hazlett, C., Imai: Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. The Annals of Applied Statistics **12**(1), 156–177 (2018)

[5] Rosenfeld, N., Mansour, Y., Yom-Tov: Predicting counterfactuals from large historical data and small randomized trials. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 602–609 (2017)

[6] Yuan, B., Hsia, J.-Y., Yang: Improving ad click prediction by considering non-displayed events. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 329–338 (2019)

[7] Lada, A., Peysakhovich, A., Aparicio: Observational data for heterogeneous treatment effects with application to recommender systems. In: Proceedings of the 2019 ACM Conference on Economics and Computation, pp. 199–213 (2019)

[8] Schnabel, T., Swaminathan, A., Singh: Recommendations as treatments: Debiasing learning and evaluation. In: International Conference on Machine Learning, pp. 1670–1679 (2016). PMLR

[9] Saito, Y.: Eliminating bias in recommender systems via pseudo-labeling. arXiv preprint arXiv:1910.01444 (2019)

[10] Swaminathan, A., Krishnamurthy, A., Agarwal: Off-policy evaluation for slate recommendation. Advances in Neural Information Processing Systems **30** (2017)

[11] Zhang, W., Bao, W., Liu: A causal perspective to unbiased conversion rate estimation on data missing not at random. arXiv preprint arXiv:1910.09337 (2019)

[12] Bonner, S., Vasile, F.: Causal embeddings for recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 104–112 (2018)

[13] Wang, X., Zhang, R., Sun: Doubly robust joint learning for recommendation on data missing not at random. In: International Conference on Machine Learning, pp. 6638–6647 (2019). PMLR

[14] Shalit, U.: Can we learn individual-level treatment policies from clinical data? Biostatistics **21**(2), 359–362 (2020)

[15] Kessler, R.C., Bossarte, R.M., Luedtke: Machine learning methods for developing precision treatment rules with observational data. Behaviour Research and Therapy **120**, 103412 (2019)

[16] Atan, O., Jordon, J., der Schaar, V.: Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[17] Atan, O., Zame, W.R., van der Schaar: Learning optimal policies from observational data. arXiv preprint arXiv:1802.08679 (2018)

[18] Kallus, N., Uehara, M.: Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. Advances in Neural Information Processing Systems **32** (2019)

[19] Tennenholtz, G., Shalit, U., Mannor: Off-policy evaluation in partially observable environments. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10276–10283 (2020)

[20] Zou, H., Kuang, K., Chen: Focused context balancing for robust offline policy evaluation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 696–704 (2019)

[21] Kallus, N., Zhou, A.: Confounding-robust policy improvement. Advances in Neural Information Processing Systems **31** (2018)

[22] Pryzant, R., Shen, K., Jurafsky: Deconfounded lexicon induction for interpretable social science. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1615–1625 (2018)

[23] Egami, N., Fong, C.J., Grimmer: How to make causal inferences using texts. arXiv preprint arXiv:1802.02163 (2018)

[24] Wood-Doughty, Z., Shpitser, I., Dredze: Challenges of using text classifiers for causal inference. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2018, p. 4586 (2018). NIH Public Access

[25] Keith, K.A., Jensen, D., O'Connor: Text and causal inference: A review of using text to remove confounding from causal estimates. arXiv preprint arXiv:2005.00649 (2020)

[26] Yao, L., Li, S., Li: On the estimation of treatment effect with text covariates. In: International Joint Conference on Artificial Intelligence (2019)

[27] Niu, Y., Tang, K., Zhang: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12700–12710 (2021)

[28] Ray, A., Sikka, K., Divakaran: Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. arXiv preprint arXiv:1909.04696 (2019)

[29] Shah, M., Chen, X., Rohrbach: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6649–6658 (2019)

[30] Agarwal, V., Shetty, R., Fritz: Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9690–9698 (2020)

[31] Wang, T., Huang, J., Zhang: Visual commonsense representation learning via causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 378–379 (2020)

[32] Zhao, S., Heffernan, N.: Estimating individual treatment effect from educational studies with residual counterfactual networks. International Educational Data Mining Society (2017)

[33] Johansson, F., Shalit, U., Sontag: Learning representations for counterfactual inference. In: International Conference on Machine Learning, pp. 3020–3029 (2016). PMLR

[34] Schwab, P., Linhardt, L., Karlen: Perfect match: A simple method for learning representations for counterfactual inference with neural networks. arXiv preprint arXiv:1810.00656 (2018)

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

[36] Yao, L., Li, S., Li: Representation learning for treatment effect estimation from observational data. Advances in Neural Information Processing Systems **31** (2018)

[37] Yoon, J., Jordon, J., Schaar, V.D.: Ganite: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations (2018)

[38] Gheisari, M., Wang, G., Bhuiyan: A survey on deep learning in big data. In: 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 2, pp. 173–180 (2017). IEEE

[39] França, R.P., Monteiro, A.C.B., Arthur: An overview of deep learning in big data, image, and signal processing in the modern digital age. Trends in Deep Learning Methodologies, 63–87 (2021)

[40] Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. Information Fusion **42**, 146–157 (2018)

[41] Jan, B., Farman, H., Khan: Deep learning in big data analytics: a comparative study. Computers & Electrical Engineering **75**, 275–287 (2019)

[42] Chakraborty, S., Tomsett, R., Raghavendra: Interpretability of deep learning models: A survey of results. In: 2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), pp. 1–6 (2017). IEEE

[43] Zhang, Q.-s., Zhu, S.-C.: Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering **19**(1), 27–39 (2018)

[44] Linardatos, P., Papastefanopoulos, V., Kotsiantis: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)

[45] Louizos, C., Shalit, U., Mooij: Causal effect inference with deep latent-variable models. Advances in Neural Information Processing systems **30** (2017)

[46] Yao, L., Li, S., Li: Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 1432–1437 (2019). IEEE

[47] Zhang, Y., Bellot, A., Schaar: Learning overlapping representations for the estimation of individualized treatment effects. In: International Conference on Artificial Intelligence and Statistics, pp. 1005–1014 (2020). PMLR

[48] Schwab, P., Linhardt, L., Bauer: Learning counterfactual representations for estimating individual dose-response curves. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 5612–5619 (2020)

[49] Nie, L., Ye, M., Liu: Vcnet and functional targeted regularization for learning causal effects of continuous treatments. arXiv preprint arXiv:2103.07861 (2021)

[50] Altman, N., Krzywinski, M.: Points of significance: Association, correlation and causation. Nature Methods **12**(10) (2015)

[51] Nauta, M., Bucur, D., Seifert: Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction **1**(1), 19 (2019)

[52] Glymour, C., Zhang, K., Spirtes: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10**, 524 (2019)

[53] Vowels, M.J., Camgoz, N.C., Bowden: D'ya like dags? a survey on structure learning and causal discovery. ACM Computing Surveys (CSUR) (2021)

[54] Hammerton, G., Munafò, M.R.: Causal inference with observational data: the need for triangulation of evidence. Psychological Medicine **51**(4), 563–578 (2021)

[55] Pawlowski, N., Coelho de Castro, D., Glocker: Deep structural causal models for tractable counterfactual inference. Advances in Neural Information Processing Systems **33**, 857–869 (2020)

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

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

[58] Wu, A., Kuang, K., Yuan: Learning decomposed representation for counterfactual inference. arXiv preprint arXiv:2006.07040 (2020)

[59] Shi, C., Blei, D., Veitch, V.: Adapting neural networks for the estimation of treatment effects. Advances in Neural Information Processing Systems **32** (2019)

[60] Kim, H., Shin, S., Jang: Counterfactual fairness with disentangled causal effect variational autoencoder. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8128–8136 (2021)

[61] Lee, C., Mastronarde, N., van der Schaar: Estimation of individual treatment effect in latent confounder models via adversarial learning. arXiv preprint arXiv:1811.08943 (2018)

[62] Bica, I., Jordon, J., van der Schaar: Estimating the effects of continuous-valued interventions using generative adversarial networks. Advances in Neural Information Processing Systems **33**, 16434–16445 (2020)

[63] Sauer, A., Geiger, A.: Counterfactual generative networks. arXiv preprint arXiv:2101.06046 (2021)

[64] Lim, B.: Forecasting treatment responses over time using recurrent marginal structural networks. Advances in Neural Information Processing Systems **31** (2018)

[65] Yao, L., Li, Y., Li: Sci: Subspace learning based counterfactual inference for individual treatment effect estimation. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 3583–3587 (2021)

[66] Zhang, Y.-F., Zhang, H., Lipton: Can transformers be strong treatment effect estimators? arXiv preprint arXiv:2202.01336 (2022)

[67] Zhu, L., Lu, W., Song: Causal effect estimation and optimal dose suggestions in mobile health. In: International Conference on Machine Learning, pp. 11588–11598 (2020). PMLR

[68] Muentener, P., Bonawitz, E.: The development of causal reasoning (2018)

[69] Schölkopf, B.: Causality for machine learning. In: Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 765–804 (2022)

[70] Kuang, K., Li, L., Geng: Causal inference. Engineering **6**(3), 253–263 (2020)

[71] Schölkopf, B., Locatello, F., Bauer: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)

[72] Sanchez, P., Voisey, J.P., Xia: Causal machine learning for healthcare and precision medicine. Royal Society Open Science **9**(8), 220638 (2022)

[73] Chen, H., Du, K., Yang, X., Li, C.: A Review and Roadmap of Deep Learning Causal Discovery in Different Variable Paradigms (2022)

[74] Gao, C., Zheng, Y., Wang, W., Feng, F.: Causal Inference in Recommender Systems: A Survey and Future Directions (2022)

[75] Kaddour, J., Lynch, A., Liu: Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475 (2022)

[76] Deng, Z., Zheng, X., Tian: Deep causal learning: Representation, discovery and inference. arXiv preprint arXiv:2211.03374 (2022)

[77] Dehejia, R.H., Wahba: Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association **94**(448), 1053–1062 (1999)

[78] Brooks-Gunn, J., Liaw: Effects of early intervention on cognitive function of low birth weight preterm infants. The Journal of Pediatrics **120**(3), 350–359 (1992)

[79] LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604–620 (1986)

[80] Saini, S.K., Dhamnani, S., Ibrahim: Multiple treatment effect estimation using deep generative model with task embedding. In: The World Wide Web Conference, pp. 1601–1611 (2019)

[81] Qian, Z., Alaa, A.M., van der Schaar: Between-centre differences for covid-19 icu mortality from early data in england. Intensive Care Medicine

46(9), 1779–1780 (2020)

[82] Haase, N., Plovsing, R., Christensen: Characteristics, interventions, and longer term outcomes of covid-19 icu patients in denmark—a nationwide, observational study. Acta Anaesthesiologica Scandinavica **65**(1), 68–75 (2021)

[83] Weinstein, J.N., Collisson, E.A., Mills: The cancer genome atlas pan-cancer analysis project. Nature Genetics **45**(10), 1113–1120 (2013)

[84] Alaa, A.M., Weisz, M., Schaar, V.D.: Deep counterfactual networks with propensity-dropout. arXiv preprint arXiv:1706.05966 (2017)

[85] Johansson, F.D., Kallus, N., Shalit: Learning weighted representations for generalization across designs. arXiv preprint arXiv:1802.08598 (2018)

[86] Hassanpour, N., Greiner, R.: Counterfactual regression with importance sampling weights. In: IJCAI, pp. 5880–5887 (2019)

[87] Zhang, Z., Lan, Q., Ding: Reducing selection bias in counterfactual reasoning for individual treatment effects estimation. arXiv preprint arXiv:1912.09040 (2019)

[88] Hassanpour, N., Greiner, R.: Learning disentangled representations for counterfactual regression. In: International Conference on Learning Representations (2019)

[89] Li, Y., Kuang, K., Li: Continuous treatment effect estimation via generative adversarial de-confounding. In: Proceedings of the 2020 KDD Workshop on Causal Discovery, pp. 4–22 (2020). PMLR

[90] Bica, I., Jordon, J., van der Schaar: Individualised dose-response estimation using generative adversarial nets. ICLR 2020 Conference Blind Submission (2019)

[91] Bica, I., Alaa, A.M., Jordon: Estimating counterfactual treatment outcomes over time through adversarially balanced representations. arXiv preprint arXiv:2002.04083 (2020)

[92] Bica, I., Alaa, A., Schaar, V.D.: Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In: International Conference on Machine Learning, pp. 884–895 (2020). PMLR

[93] Du, X., Sun, L., Duivesteijn: Adversarial balancing-based representation learning for causal effect inference with observational data. Data Mining and Knowledge Discovery **35**(4), 1713–1738 (2021)

[94] Assaad, S., Zeng, S., Tao: Counterfactual representation learning with balancing weights. In: International Conference on Artificial Intelligence and Statistics, pp. 1972–1980 (2021). PMLR

[95] Sun, X., Wu, B., Zheng: Latent causal invariant model. arXiv preprint arXiv:2011.02203 (2020)

[96] Zou, H., Cui, P., Li: Counterfactual prediction for bundle treatment. Advances in Neural Information Processing Systems **33**, 19705–19715 (2020)

[97] Parbhoo, S., Bauer, S., Schwab: Ncore: Neural counterfactual representation learning for combinations of treatments. arXiv preprint arXiv:2103.11175 (2021)

[98] Guo, Z., Zheng, S., Liu: Cetransformer: Casual effect estimation via transformer based representation learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 524–535 (2021). Springer

[99] Hatt, T., Feuerriegel, S.: Estimating average treatment effects via orthogonal regularization. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 680–689 (2021)

[100] Curth, A., van der Schaar, M.: On inductive biases for heterogeneous treatment effect estimation. Advances in Neural Information Processing Systems **34** (2021)

[101] Qian, Z., Curth, A., van der Schaar: Estimating multi-cause treatment effects via single-cause perturbation. Advances in Neural Information Processing Systems **34** (2021)

[102] Qian, Z., Zhang, Y., Bica: Synctwin: Treatment effect estimation with longitudinal outcomes. Advances in Neural Information Processing Systems **34** (2021)

[103] Zhong, K., Xiao, F., Ren: Descn: Deep entire space cross networks for individual treatment effect estimation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4612–4620 (2022)

[104] Zhou, G., Yao, L., Xu, X., Wang, C., Zhu, L.: Cycle-balanced representation learning for counterfactual inference. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), pp. 442–450 (2022). SIAM

[105] Li, X., Yao, L.: Contrastive individual treatment effects estimation. In: 2022 IEEE International Conference on Data Mining (ICDM), pp. 1053–1058 (2022). IEEE

[106] Wang, X., Lyu, S., Wu, X., Wu, T., Chen, H.: Generalization bounds for estimating causal effects of continuous treatments. In: Advances in Neural Information Processing Systems

[107] Melnychuk, V., Frauen, D., Feuerriegel, S.: Causal transformer for estimating counterfactual outcomes. arXiv preprint arXiv:2204.07258 (2022)

[108] Liu, R., Chen, P.-Y., Zhang: Cure: A pre-training framework on large-scale patient data for treatment effect estimation. medRxiv (2022)

[109] Reddy, A.G., Balasubramanian, V.N.: Estimating treatment effects using neurosymbolic program synthesis. arXiv preprint arXiv:2211.04370 (2022)

[110] Liu, R., Yin, C., Zhang, P.: Estimating individual treatment effects with time-varying confounders. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 382–391 (2020). IEEE

[111] Guo, R., Li, J., Liu, H.: Learning individual causal effects from networked observational data. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 232–240 (2020)

[112] Berrevoets, J., Jordon, J., Bica: Organite: Optimal transplant donor organ offering using an individual treatment effect. Advances in Neural Information Processing Systems **33**, 20037–20050 (2020)

[113] Ma, J., Guo, R., Zhang: Multi-cause effect estimation with disentangled confounder representation. In: IJCAI, pp. 2790–2796 (2021)

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

[115] Curth, A., van der Schaar, M.: Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In: International Conference on Artificial Intelligence and Statistics, pp. 1810–1818 (2021). PMLR

[116] Moraffah, R., Sheth, P., Karami: Causal inference for time series analysis: Problems, methods and evaluation. Knowledge and Information Systems, 1–45 (2021)

[117] Chu, Z., Rathbun, S.L., Li: Learning infomax and domain-independent representations for causal effect inference with real-world data. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), pp. 433–441 (2022). SIAM

[118] Qidong, L., Feng, T., Weihua: A new representation learning method for individual treatment effect estimation: Split covariate representation network. In: Asian Conference on Machine Learning, pp. 811–822 (2020). PMLR

[119] Prichard, B., Gillam, P.: Assessment of propranolol in angina pectoris. clinical dose response curve and effect on electrocardiogram at rest and on exercise. British heart journal **33**(4), 473 (1971)

[120] Schneider, A.B., Ron, E., Lubin: Dose-response relationships for radiation-induced thyroid cancer and thyroid nodules: evidence for the prolonged effects of radiation on the thyroid. The Journal of Clinical Endocrinology & Metabolism **77**(2), 362–369 (1993)

[121] Threlfall, T.J., English, D.R.: Sun exposure and pterygium of the eye: a dose-response curve. American Journal of Ophthalmology **128**(3), 280–287 (1999)

[122] Zhou, G., Yao, L., Xu, X., Wang, C., Zhu, L.: Learning to infer counterfactuals: Meta-learning for estimating multiple imbalanced treatment effects. arXiv preprint arXiv:2208.06748 (2022)

[123] Zou, H., Li, B., Han: Counterfactual prediction for outcome-oriented treatments. In: International Conference on Machine Learning, pp. 27693–27706 (2022). PMLR

[124] Chen, X., Liu, Z., Yu: Adversarial learning for incentive optimization in mobile payment marketing. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2940–2944 (2021)

[125] Mondal, A., Majumder, A., Chaoji: Memento: Neural model for estimating individual treatment effects for multiple treatments. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 3381–3390 (2022)

[126] Lopez, R., Li, C., Yan: Cost-effective incentive allocation via structured counterfactual inference. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 4997–5004 (2020)

[127] Liu, Z., Wang, D., Yu: Graph representation learning for merchant incentive optimization in mobile payment marketing. In: Proceedings of the

28th ACM International Conference on Information and Knowledge Management, pp. 2577–2584 (2019)

[128] Yu, L., Wu, Z., Cai: Joint incentive optimization of customer and merchant in mobile payment marketing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15000–15007 (2021)

[129] Zhang, Y., Wang, D., Li: User retention: A causal approach with triple task modeling. In: IJCAI, pp. 3399–3405 (2021)

[130] Chen, J., Dong, H., Wang: Bias and debias in recommender system: A survey and future directions. arXiv preprint arXiv:2010.03240 (2020)

[131] Liu, D., Cheng, P., Zhu: Mitigating confounding bias in recommendation via information bottleneck. In: Fifteenth ACM Conference on Recommender Systems, pp. 351–360 (2021)

[132] Wang, H., Chang, T.-W., Liu: Escm: Entire space counterfactual multi-task model for post-click conversion rate estimation. arXiv preprint arXiv:2204.05125 (2022)

[133] Liu, D., Cheng, P., Dong: A general knowledge distillation framework for counterfactual recommendation via uniform data. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 831–840 (2020)

[134] Chen, J., Dong, H., Qiu: Autodebias: Learning to debias for recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–30 (2021)

[135] Chemla, G., Hennessy, C.: Signaling, random assignment, and causal effect estimation (2020)

[136] Tiffin, M.A.J.: Machine Learning and Causality: The Impact of Financial Crises on Growth. International Monetary Fund, ??? (2019)

[137] Athey, S., Wager, S.: Efficient policy learning. Technical report (2017)

[138] Arpino, B., Mattei, A.: Assessing the causal effects of financial aids to firms in tuscany allowing for interference. The Annals of Applied Statistics **10**(3), 1170–1194 (2016)

[139] Zheng, X., Dan, C., Aragam: Learning sparse nonparametric dags. In: International Conference on Artificial Intelligence and Statistics, pp. 3414–3425 (2020). PMLR

[140] Tanimoto, A., Sakai, T., Takenouchi: Regret minimization for causal inference on large treatment space. In: International Conference on Artificial

Intelligence and Statistics, pp. 946–954 (2021). PMLR

[141] Assaad, C.K., Devijver, E., Gaussier: Survey and evaluation of causal discovery methods for time series. Journal of Artificial Intelligence Research **73**, 767–819 (2022)

[142] Harada, S., Kashima, H.: Graphite: Estimating individual effects of graph-structured treatments. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 659–668 (2021)

[143] Seedat, N., Imrie, F., Bellot: Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. arXiv preprint arXiv:2206.08311 (2022)

[144] Alaa, A., Van Der Schaar, M.: Validating causal inference models via influence functions. In: International Conference on Machine Learning, pp. 191–201 (2019). PMLR

[145] Hünermund, P., Bareinboim, E.: Causal inference and data fusion in econometrics. arXiv preprint arXiv:1912.09104 (2019)

[146] Morck, R., Yeung, B.: Economics, history, and causation. Business History Review **85**(1), 39–63 (2011)

[147] Athey, S., Imbens, G.W.: Machine learning methods that economists should know about. Annual Review of Economics **11**, 685–725 (2019)

[148] Finkelstein, A., Hendren, N.: Welfare analysis meets causal inference. Journal of Economic Perspectives **34**(4), 146–67 (2020)

[149] Angrist, J.D., Pischke, J.-S.: The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of Economic Perspectives **24**(2), 3–30 (2010)

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

[151] Belloni, A., Chernozhukov, V., Hansen: Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies **81**(2), 608–650 (2014)

[152] Angrist, J.: Empirical strategies in economics: Illuminating the path from cause to effect. Technical report, National Bureau of Economic Research (2022)

[153] Varian, H.R.: Causal inference in economics and marketing. Proceedings of the National Academy of Sciences **113**(27), 7310–7315 (2016)

[154] Ramachandra, V.: Deep learning for causal inference. arXiv preprint arXiv:1803.00149 (2018)

[155] Webbink, D.: Causal effects in education. Journal of Economic Surveys **19**(4), 535–560 (2005)

[156] Card, D.: The causal effect of education on earnings. Handbook of Labor Economics **3**, 1801–1863 (1999)

[157] Schlotter, M., Schwerdt, G., Woessmann: Econometric methods for causal evaluation of education policies and practices: a non-technical guide. Education Economics **19**(2), 109–137 (2011)

[158] Sales, A.C., Wilks, A., Pane: Student usage predicts treatment effect heterogeneity in the cognitive tutor algebra i program. International Educational Data Mining Society (2016)

[159] Sales, A., Prihar, E., Heffernan, N., Pane: The effect of an intelligent tutor on performance on specific posttest problems. International Educational Data Mining Society (2021)

[160] de Carvalho, W.F., Couto, B.R.G.M., Ladeira: Applying causal inference in educational data mining: A pilot study. In: CSEDU (1), pp. 454–460 (2018)

[161] Chen, L.K., Ramsey, J., Dubrawski, A.: Affect, support, and personal factors: Multimodal causal models of one-on-one coaching. Journal of Educational Data Mining **13**(3), 36–68 (2021)

[162] Ramsey, J.D., Zhang, K., Glymour: Tetrad—a toolbox for causal discovery. In: 8th International Workshop on Climate Informatics (2018)

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

[164] Almond, D., Chay, K.Y., Lee: The costs of low birth weight. The Quarterly Journal of Economics **120**(3), 1031–1083 (2005)

[165] Dorie, V., Hill, J., Shalit: Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statistical Science **34**(1), 43–68 (2019)

[166] Geng, C., Paganetti, H., Grassberger: Prediction of treatment response for combined chemo-and radiation therapy for non-small cell lung cancer patients using a bio-mathematical model. Scientific Reports **7**(1), 1–12 (2017)

[167] Barbolosi, D., Iliadis, A.: Optimizing drug regimens in cancer chemotherapy: a simulation study using a pk–pd model. Computers in Biology and Medicine **31**(3), 157–172 (2001)

[168] Eigenmann, M.J., Frances, N., Lavé: Pkpd modeling of acquired resistance to anti-cancer drug treatment. Journal of Pharmacokinetics and Pharmacodynamics **44**(6), 617–630 (2017)

[169] Johnson, A.E., Pollard: Mimic-iii, a freely accessible critical care database. Scientific Data **3**(1), 1–9 (2016)

[170] Saeed, M., Villarroel, M., Reisner: Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical Care Medicine **39**(5), 952 (2011)

[171] Jack Jr, C.R., Bernstein, M.A., Fox: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)

[172] Humpel, C., Hochstrasser, T.: Cerebrospinal fluid and blood biomarkers in alzheimer's disease. World Journal of Psychiatry **1**(1), 8 (2011)

[173] Ramani, C., Davis, E.M., Kim: Post-icu covid-19 outcomes: a case series. Chest **159**(1), 215–218 (2021)

[174] Herrett, E., Gallagher, A.M., Bhaskaran: Data resource profile: clinical practice research datalink (cprd). International Journal of Epidemiology **44**(3), 827–836 (2015)

[175] Bisgin, H., Agarwal, N., Xu: Does similarity breed connection?-an investigation in blogcatalog and last. fm communities. In: 2010 IEEE Second International Conference on Social Computing, pp. 570–575 (2010). IEEE

[176] Stephens, M.: Flickr. Library Technology Reports **42**(4), 58–62 (2008)

[177] Lee, K., Bargagli-Stoffi, F.J., Dominici: Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. arXiv preprint arXiv:2009.09036 (2020)

[178] Bargagli-Stoffi, F.J., De-Witte, K., Gnecco: Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach. arXiv preprint arXiv:1905.12707 (2019)

[179] Imai, K., Li, M.L.: Experimental evaluation of individualized treatment rules. Journal of the American Statistical Association, 1–15 (2021)

[180] Curth, A., Svensson, D., Weatherall: Really doing great at estimating

cate? a critical look at ml benchmarking practices in treatment effect estimation. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[181] Kaddour, J., Zhu, Y., Liu: Causal effect inference for structured treatments. Advances in Neural Information Processing Systems **34**, 24841–24854 (2021)

[182] Young, J.D., Andrews, B., Cooper: Learning latent causal structures with a redundant input neural network. In: Proceedings of the 2020 KDD Workshop on Causal Discovery, pp. 62–91 (2020). PMLR

[183] Zhang, J., Pham, V.V.H., Liu: Identifying mirna synergism using multiple-intervention causal inference. BMC bioinformatics **20**(23), 1–11 (2019)

[184] Schmidt, C., Huegle, J., Horschig: Out-of-core gpu-accelerated causal structure learning. In: International Conference on Algorithms and Architectures for Parallel Processing, pp. 89–104 (2019). Springer

[185] Young, J.: Deep learning for causal structure learning applied to cancer pathway discovery. PhD thesis, University of Pittsburgh (2020)

[186] Chou, C.-Y., Chang, W.-I., Horng: Numerical modeling of nanodrug distribution in tumors with heterogeneous vasculature. PLoS One **12**(12), 0189802 (2017)

[187] Tu, R., Zhang, K., Bertilson: Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. Advances in Neural Information Processing Systems **32** (2019)

[188] Patmanidis, S., Charalampidis, A.C., Kordonis: Individualized growth prediction of mice skin tumors with maximum likelihood estimators. Computer Methods and Programs in Biomedicine **185**, 105165 (2020)

[189] Bica, I., Alaa, A.M., Lambert: From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clinical Pharmacology & Therapeutics **109**(1), 87–100 (2021)

[190] Karboub, K., Tabaa, M.: A machine learning based discharge prediction of cardiovascular diseases patients in intensive care units. In: Healthcare, vol. 10, p. 966 (2022). Multidisciplinary Digital Publishing Institute

[191] Liu, C., Sun, X., Wang: Learning causal semantic representation for out-of-distribution prediction. Advances in Neural Information Processing Systems **34**, 6155–6170 (2021)

[192] Sun, X., Wu, B., Zheng: Recovering latent causal factor for generalization to distributional shifts. Advances in Neural Information Processing Systems **34**, 16846–16859 (2021)

[193] Wang, T., Zhou, C., Sun: Causal attention for unbiased visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3091–3100 (2021)

[194] Yuan, Z., Peng, X., Wu: Meta-learning causal feature selection for stable prediction. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2021). IEEE

[195] Wang, Y., Li, X., Qi: Meta-causal feature learning for out-of-distribution generalization. arXiv preprint arXiv:2208.10156 (2022)

[196] Pölsterl, S., Wachinger, C.: Estimation of causal effects in the presence of unobserved confounding in the alzheimer's continuum. In: International Conference on Information Processing in Medical Imaging, pp. 45–57 (2021). Springer

[197] Zhang, B., Guo, X., Lin: Counterfactual inference graph network for disease prediction. Knowledge-Based Systems, 109722 (2022)

[198] Wachinger, C., Becker, B.G., Rieckmann: Quantifying confounding bias in neuroimaging datasets with causal inference. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 484–492 (2019). Springer

[199] Wang, R., Chaudhari, P., Davatzikos: Harmonization with flow-based causal inference. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 181–190 (2021). Springer

[200] Van Goethem, N., Serrien, B., Vandromme: Conceptual causal framework to assess the effect of sars-cov-2 variants on covid-19 disease severity among hospitalized patients. Archives of Public Health **79**(1), 1–12 (2021)

[201] Wilde, H., Mellan, T., Hawryluk: The association between mechanical ventilator compatible bed occupancy and mortality risk in intensive care patients with covid-19: a national retrospective cohort study. BMC Medicine **19**(1), 1–12 (2021)

[202] Mastakouri, A., Schölkopf, B.: Causal analysis of covid-19 spread in germany. Advances in Neural Information Processing Systems **33**, 3153–3163 (2020)

[203] Gvozdenović, E., Malvisi, L., Cinconze: Causal inference concepts applied to three observational studies in the context of vaccine development: from

theory to practice. BMC Medical Research Methodology **21**(1), 1–10 (2021)

[204] Charpignon, M.-L., Vakulenko-Lagun, B., Zheng: Drug repurposing of metformin for alzheimer's disease: Combining causal inference in medical records data and systems pharmacology for biomarker identification. MedRxiv (2021)

[205] Rao, S., Mamouei, M., Salimi-Khorshidi: Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records. arXiv preprint arXiv:2202.03487 (2022)

[206] Zhu, J., Gallego, B.: Cds–causal inference with deep survival model and time-varying covariates. arXiv preprint arXiv:2101.10643 (2021)

[207] Coulombe, J.: Causal inference on the marginal effect of an exposure: Addressing biases due to covariate-driven monitoring times and confounders. PhD thesis, McGill University (Canada) (2021)

[208] Li, J., Guo, R., Liu: Adaptive unsupervised feature selection on attributed networks. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 92–100 (2019)

[209] Li, J., Hu, X., Tang: Unsupervised streaming feature selection in social media. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1041–1050 (2015)

[210] Wu, A., Yuan, J., Kuang: Learning decomposed representations for treatment effect estimation. IEEE Transactions on Knowledge and Data Engineering (2022)

[211] Dehejia, R.H., Wahba: Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics **84**(1), 151–161 (2002)

[212] Smith, J.A., Todd, P.E.: Does matching overcome lalonde's critique of nonexperimental estimators? Journal of Econometrics **125**(1-2), 305–353 (2005)