A SURVEY OF DEEP CAUSAL MODELS

Zongyu Li
Beijing Jiaotong University, Beijing, China
Beijing Key Laboratory of Advanced Information Science and Network Technology,
Beijing, China
zongyuli@bjtu.edu.cn

ZhenFeng Zhu*
Beijing Jiaotong University, Beijing, China
Beijing Key Laboratory of Advanced Information Science and Network Technology,
Beijing, China
zhfzhu@bjtu.edu.cn

ABSTRACT

The concept of causality plays an important role in human cognition . In the past few decades, causal inference has been well developed in many fields, such as computer science, medicine, economics, and education. With the advancement of deep learning techniques, it has been increasingly used in causal inference against counterfactual data. Typically, deep causal models map the characteristics of covariates to a representation space and then design various objective optimization functions to estimate counterfactual data unbiasedly based on the different optimization methods. This paper focuses on the survey of the deep causal models, and its core contributions are as follows: 1) we provide relevant metrics under multiple treatments and continuous-dose treatment; 2) we incorporate a comprehensive overview of deep causal models from both temporal development and method classification perspectives; 3) we assist a detailed and comprehensive classification and analysis of relevant datasets and source code.

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 inference 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, estimating causal effects of observational data in advertising[3, 4, 5, 6, 7, 8, 9], developing recommender systems that are highly correlated with causal treatment effect estimates[10, 11, 12, 13, 14, 15, 16], learning optimal treatment rules for patients in medicine[17, 18, 19], estimation of ITE in reinforcement learning[20, 21, 22, 23, 24, 25, 26, 27, 28], causal inference tasks in natural language processing[29, 30, 31, 32, 33, 34], emerging computer vision and language interaction tasks[35, 36, 37, 38, 39], education[40], policy decisions[41, 42, 43, 44, 45] and improved machine learning methods[46], etc.

Deep learning contributes to the development of artificial intelligence when applied to big data[47, 48, 49, 50]. In comparison with traditional machine learning algorithms, deep learning models are more computationally efficient, more accurate, and hold good 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[51, 52, 53]. In recent years, deep learning models have been widely used for mining data for causality rather than correlation[41, 43]. Thus, deep causal models have become a core method for estimating treatment effects based on unbiased estimates[19, 44, 45, 54]. At present, many works in the field of causal inference utilize deep causal models to select reasonable treatment options[55, 56, 57, 58].

*Corresponding author

With big data, all trend variables are correlated[59], so discovering causal relationships is a challenging problem[60, 61, 62]. In terms of statistical theory, it is the most effective way to conduct **randomized controlled trials(RCT)**[63] 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. Research studies involving RCTs require a large number of samples with little variation in characteristics, which is difficult to interpret and involves ethical challenges. As a matter of fact, it is not wise to select subjects to try a drug or vaccine[64, 65]. Therefore, causal effects are usually measured directly using observational data. A central question for obtaining counterfactual results is how to deal with observational data[66]. When observational data are analyzed, treatments are not randomly assigned and the performance of samples after treatment is significantly different from the performance of ordinary samples[41, 43]. Unfortunately, we cannot observe alternative outcomes in theory since we cannot observe counterfactual results[67].

A long-standing feature of mainstream research has been the use of the potential outcome framework as a means of solving the problem of causal inference from observational data[68]. The potential outcome framework is also known as the Rubin Causal Model[69]. Causal inference is closely connected to deep learning since it is conceptualized using Rubin Causal Model. In order to enhance the accuracy and unbiasedness of estimates, several researchers have tried combining deep networks and causal models. To illustrate, consider representations of distribution balance methods[41, 43, 44], the effects of covariates confounding learning methods[54, 70, 71], methods based on generative adversarial networks[45, 72, 73] and so forth[58, 34, 74]. As deep learning methods facilitate causal inference, causal inference also contributes to the development of deep learning methods. In addition to improving the accuracy of causal effect estimation, studies of deep networks provide a plausible basis for developing deep learning algorithms[75, 76].

Various perspectives have been discussed in recent years regarding causal inference[77, 1, 78, 79, 80, 81, 82, 83, 2]. In Table 1, the titles and main points of the relevant reviews are listed. An in-depth analysis of the origins and variable development of causal inference is provided in review[77], as well as the implications of causal learning for the development of causal inference. Aside from that, an overview of traditional and cutting-edge causal learning methods, and a comparison between machine learning and causal learning, can be found in survey[1]. Many scholars have discussed how machine learning can be interpreted. Immediately afterward, to create explainable artificial intelligence algorithms, survey[79] combines causal reasoning and machine learning. As a novel perspective, causal representation learning is flourishing, and review[80] uses it to uncover high-level causal variables from low-level observations, strengthening the link between machine learning and causal inference. Due to causal machine learning's popularity in recent years, a detailed discussion of the relevance of graphical causal inference to machine learning is provided in review[78]. Furthermore, in survey[81], the author examines how recent advances in machine learning can be applied to causal inference, and provides a comprehensive interpretation of how causal machine learning can contribute to the advancement of medical science. As review[82] argues, causal discovery methods can be improved and sorted out based on deep learning, and variable paradigms can be explored to help think about and explore causal discovery methods. Causal inference in recommender systems is the focus of survey[83] , which explains how to use causal inference to extract causal relationships in order to enhance recommender systems. It has long been the potential outcome framework of statistics that bridges causal inference with deep learning, as a starting point, survey[2] examines and compares traditional statistical algorithms and machine learning algorithms for different categories that satisfy these assumptions. In light of the rapid development of deep learning algorithms, the existing literature does not take deep causal models into account when examining generalization. Therefore, from the perspective of deep network, we summarize the deep causal model in terms of time and classification. This survey provides a comprehensive review and analysis of deep causal models in recent years. It makes three core contributions: 1) We incorprate relevant metrics in the case of multiple treatments as well as continuous-dose treatment. 2) We present a comprehensive overview of deep causal models from the perspective of both method classification and temporal development. 3) We provide detailed and comprehensive support in the analysis and classification of relevant datasets and source code.

Table 1: A Summary of Main Points about Related Reviews

Survey title	**Core content**
The Development of Causal Reasoning[77]	Origin and Development of Causal Inference
Causal Inference[79]	Machine Learning Interpretability of Counterfactual Causal Inference
Causality for Machine Learning[78]	Connection of Graphical Causal Inference to Machine Learning
Toward Causal Representation Learning[80]	Exploring Causal Variables in Data by Causal Representations Learning
Causal Machine Learning for Healthcare and Precision Medicine[81]	Causal Machine Learning in Healthcare
A Survey of Learning Causality with Data: Problems and Methods[1]	Relationship between Causal Learning and Machine Learning in Big Data Situations
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[82]	The application of deep learning and a variable paradigm perspective for causal discovery
Causal Inference in Recommender Systems: A Survey and Future Directions[83]	Optimize recommender systems by extracting causal relationships through causal inference
A Survey of Deep Causal Models	In-depth Causal Inference Model from Perspective of Deep Network Development

Below is an outline of the rest of the paper. As discussed in Section 2, the deep causal models are introduced, along with definitions and assumptions. In Section 3, appropriate examples and metrics are introduced, including binary treatment, multiple treatment, and continuous dose treatment. A deep causal model is demonstrated in Section 4, which includes an overview and an analysis of it. In Section 5 the methods of the deep causal models are discussed, including distribution balance methods, covariate confounding learning methods, methods based on generative adversarial networks, methods based on time series with text input and methods based on multi-treatment and continuous-dose treatment models. A list of relevant experimental guidelines follows in Section 6. A summary of the paper is presented in Section 7.

2 Preliminaries

In this section, the basic knowledge of deep causal models is introduced, including task descriptions, mathematical concepts, pertinent assumptions, examples, and metrics.

Basically, the aim of causal effect estimation is to estimate the change in outcome that will occur if a different treatment are implemented. 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 cannot 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. It refers to how to find the most effective treatment plan based on past experimental diagnosis and medical history of the patient.

Because of the widespread accumulating of data in fields such as health care[84, 85, 86], sociology science[87, 88, 89, 90], digital marketing[91, 92, 93], and machine learning[94, 95, 96, 97, 98], observational studies are becoming increasingly important. Researchers are increasingly using deep learning networks to make counterfactual estimates based on observational data, and deep causal models can aid various fields in making optimal treatment decisions.

2.1 Definitions

Here, the basic notation definitions under the potential outcome framework[69] are illustrated. According to this framework, causation 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. Below are concepts related to causal effect estimation that are benchmarked against the relevant basic definitions in the survey[2].

Definition 1 *Sample: A sample is also known as a unit, that is, an atomic study object.*

Typical samples include a person, a patient, an object, a collection of objects or people at a given time, a classroom, or a marker[99]. Samples in a population constitute units within a dataset.

Definition 2 *Treatment: A treatment describes 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 treatments can be indicated by the T ($T \in \{0, 1, 2, \ldots, T_N\}$), where $N + 1$ designates 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, evaluated 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: Counterfactual outcomes are outcomes that differ from the factual outcomes.*

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

Definition 5 *Dose: Dose refers to the amount taken continuously during a particular treatment.*

Medical treatments involving continuous dose parameters are numerous, such as (vasopressors[100]), A set of consecutive dose schemes can be donated as D_T, the factual dose for a given treatment 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 response effect of a sample after receiving different doses of an intervention over time.*

A better fit to the dose-response curve can make the model more robust and expressive in continuous dose treatments. The set of actual and counterfactual outcome responses on the dose-response curves are $Y^F(D^F, T_i)$ and $Y^{CF}(D^{CF}, T_i^{'})$.

Definition 7 *Covariates: Covariates are variables that are not affected by treatment choice.*

Generally, covariates in the medical environment refer to the patient's demographic, medical history, experimental data, and so forth, usually denoted by X. Covariates can be separated into confounding and non-confounding variables, specifically divided into three categories[70]: instrumental factor I, which primarily affects treatment T; confounding factor C, which contributes to both treatment T and outcome Y; and adjustment factor A, which determines outcome Y.

2.2 Assumptions

Having understood the basic definition of the causal model, the following three assumptions are commonly required to realize the estimation of causal treatment effect, these basic assumptions are derived from papers[2, 101].

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

Based on this assumption, there is no interaction between samples, as well as only one version of each treatment scheme. Donations can be made as $P(Y_i|T_i, T_i^{'}, X_i) = P(Y_i|T_i, X_i)$.

Assumption 2 ***Ignorability:*** *With respect to the covariate X, the treatment distribution T is independent of the potential outcomes.*

In the assumption of ignorability, there should be no unobserved confounders. $T \perp\!\!\!\perp Y(T = T_i), Y(T = T_i^{'})|X$ needs to be satisfied.

Assumption 3 ***Overlap:*** *Depending on the covariate X, each sample has a chance of receiving an intervention.*

To estimate the counterfactual treatment effect, it must be assumed that each sample can implement any treatment scheme, otherwise the overlap assumption will not be valid. The donation is $0 < P(T = T_i|X = x) < 1$ and $0 < P(T = T_i^{'}|X = x) < 1$

3 Examples and Metrics

Deep causal models utilize different metrics to address different practical issues. An analysis and description of the different performance metrics adopted for different scenario-based applications follows. As in medicine, health care, markets, job searches, social economy, and advertising when it comes to binary treatment problems, multi-treatment problems, and continuous dose treatment problems. Only the classic examples are discussed in this section. To view the detailed dataset description, please refer to Section 6. In addition to the baseline measures of the survey[2], we expend measures for multiple treatments and continue-doses treatments.

3.1 Binary treatment

As an example of binary treatment, the most famous is **Infant Health and Development Program(IHDP)**[102]. Children's quality of child care and home visits are represented as covariates, with observations based on a certain algorithmic process and a new biased subset being omitted from the model for selection bias. Similarly, the sample **Twins**[54] about twin births in the USA are also commonly used, The treatment group and the control group correspond to the weight of the twins, and the result corresponds to the mortality rate within one year of birth. The investigators set the treatment assignments themselves in order to simulate selection bias.

Using the example above, the most basic and common performance metric is **Average Treatment Effect(ATE)**, which is determined by the following[103]:

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

where $Y(T=1)$ and $Y(T=0)$ indicate the results of the treatment and control groups in the population.

The treatment effect in a sample set is called **Conditional Average Treatment Effect (CATE)**, it is calculated as follows[2]:

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

where $Y(T=1)|X=n$ and $Y(T=0)|X=n$ represent the results of the $X=n$ treatment group and the control group under the sample set, respectively. Due to the fact that different treatments have different impacts on different sets of examples, CATE is also known as heterogeneous treatment effect. It is also possible to apply ATE and CATE to multiple treatment scenarios.

Treatment effect is typically estimated as **Individual Treatment Effect (ITE)** at the individual level, which is defined as[2]:

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

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

It is also helpful to keep in mind that another evaluation metric called **Precision in Estimation of Heterogeneous(PEHE)** is used frequently. Regardless of fact or counterfactual outcomes, PEHE requires unbiased estimates, as defined as follows[45]:

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

where $Y_1^F(n)$, $Y_0^F(n)$ and $Y_1^{CF}(n)$, $Y_0^{CF}(n)$ respectively indicate unbiased estimates of fact and counterfactual for the treatment and control groups.

Another widely used research sample is **Jobs**[88, 103], which is widely used by causal researchers. A total of eight covariates are included in this study, including age, education, ethnicity, and income in 1974 and 1975 when vocational training is applied and the result is income and employment status post-training.

As an alternative to ATE and ITE, this sample can estimate treatment effects by using **Average Treatment effect on the Treated group (ATT)**. ATT is defined as[2]:

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

where $Y(T=1)|T=1$ and $Y(T=0)|T=1$ correspond to the treatment and control outcomes for treatment groups respectively.

Since only factual data is available for Jobs, the testing set comes from RCT. Performance metrics for **policy risk** $(\mathcal{R}_{pol}(\pi))$ can be described as follows[43]:

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

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

3.2 Multiple treatment

Continuing from the binary treatment scenario, this subsection discusses multiple treatment scenarios. Besides the IHDP, the **News**[42] example is used frequently for questions involving multiple treatments. As shown in this example, how news items are perceived by media consumers. There is a covariate of the number of words in a news item, and in the treatment, there are viewing tools available, such as smartphones, tablets, desktops, and TVs.

An accurate estimate of treatment effect can be determined for all subsets of the group using **Root Mean Square Error (RMSE)** rather than PEHE. RMSE is defined as[104]:

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

$Y^F(n,j)$ displays the true result when the j^{th} subset is applied to compute the i^{th} observation, whereas $Y^{CF}(n,j)$ displays the predicted result when the j^{th} subset is applied to compute the i^{th} observation. Absolute error is better measured by RMSE.

To measures the average of the RMSE between the actual and estimated difference between ITE for each treatment with no treatment ITE, we can incorporate multiple treatments into the calculation of **Average PEHE**[104], multiple treatments can be taken into account:

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

The set T represents the sample set with no treatment applied, whereas T_0 represents the sample set with no treatment applied.

It is worth mentioning that the case study on the estimation of multi-cause treatment effects of **COVID-19**[105, 106] Hospitalization in England Surveillance System(CHESS) has become a research hotspot. In this example, there is information on individual-level risk factors, treatments, and outcomes of 3090 patients admitted at the height of the outbreak, including age, multiple morbidity, ventilatory support, antiviral treatment, etc. CATE and RMSE typically serve as performance metrics in this sample.

3.3 Continuous dose treatment

The Cancer Genome Atlas (TCGA)[107] serves as a classic example of continuous dose treatment. For instance, TCGA collected gene expression data for 9659 individuals with different types of cancer. Drug therapy, chemotherapy, and surgery are the treatment options, and the outcome is a risk of cancer recurrence after treatment. Also, **Mechanical Ventilation in the Intensive Care Unit (MVICU)**[57] example can be used to illustrate continuous dose treatment. An example of response to mechanical ventilation configurations in the intensive care unit is included here. The example is taken from a publicly available database, **MIMIC III**[108], which contains comprehensive and detailed clinical information on a large and diverse group of ICU inpatients. An indicator of covariance is the last measurement of various biological signals, including respiratory, cardiac, and ventilation signals. One of the clinical criteria for the diagnosis of Acute Respiratory Distress Syndrome (ARDS)[109] is arterial blood gas readings of arterial oxygen partial pressure to fractional inspired oxygen.

For continuous dose treatment, the sample dose-response curve is accepted as a measure. On the other hand, the metrics on the test set are different[73]. In terms of **Mean Integral Squared Error (MISE)**, the model measures how accurately it estimates patient outcomes over the dose space, which is defined as[73]:

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

Treatment T is representative of the set of treatments in the space, while sample K is the number of samples. For a given treatment, T_n lies within the dose space D_{T_n}. A pain score of $Y_n(T_n,u)$ is equivalent to a model-determined outcome and $\hat{Y_n}(T_n,u)$ is equivalent to the optimal treatment dose, respectively.

As well as this, the **Mean Dose Policy Error (DPE)** is another good measure of a model's ability to predict the optimal dose point for each individual treatment, and it can be defined as[73]:

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

Specifically, $D^*_{T_n}$ and $\hat{D}^*_{T_n}$ represent the true optimal dose and the model-determined optimal dose under a treatment, respectively. With SciPy's Sequential Least Squares Point, 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)** needs to be calculated. PE is defined as[73]:

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

In the case, 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 is selected.

It is also possible, after simulation data synthesis and expansion, to apply many binary examples to multi-treatment or continuous-dose situations, such as ihdp, twins, and so on. Meanwhile, binary estimation can also be performed using examples such as News and ACIC. Detailed descriptions of the related sample datasets are given in Section 6.

4 Development

With a solid understanding of the background and basic definitions, this section moves into the core of the deep causal model. An overview of deep causal models and their development over the past six years are provided here, including an analysis of 41 deep causal models based on the timeline.

4.1 Overview

The study of deep causal models has become increasingly popular in the last few years. As deep learning has advanced, various deep causal models have become more accurate and efficient at estimating causal effects. According to Figure 1, about 40 classic deep causal models from June 2016 to February 2022 are listed, including their detailed names and when they are proposed.

Deep causal models have been developed since 2016. For the first time, Johansson et al. publish **Learning Representations for Counterfactual Inference**[41], and propose the algorithm framework BNN and BLR[41], which combines deep learning with the causal effect estimation problem, and transforms the causal inference problem into a domain adaptation problem. A number of models, including DCN-PD[110], TARNet and CFRNet[43], have been proposed since then. In this regard, it is important to note that the CEVAE[54] model proposed by Louizos et al. in December 2017, based on deep network classical structural parameter autoencoders VAE, focuses on confounding factors and their impact on the estimation of causal effects.

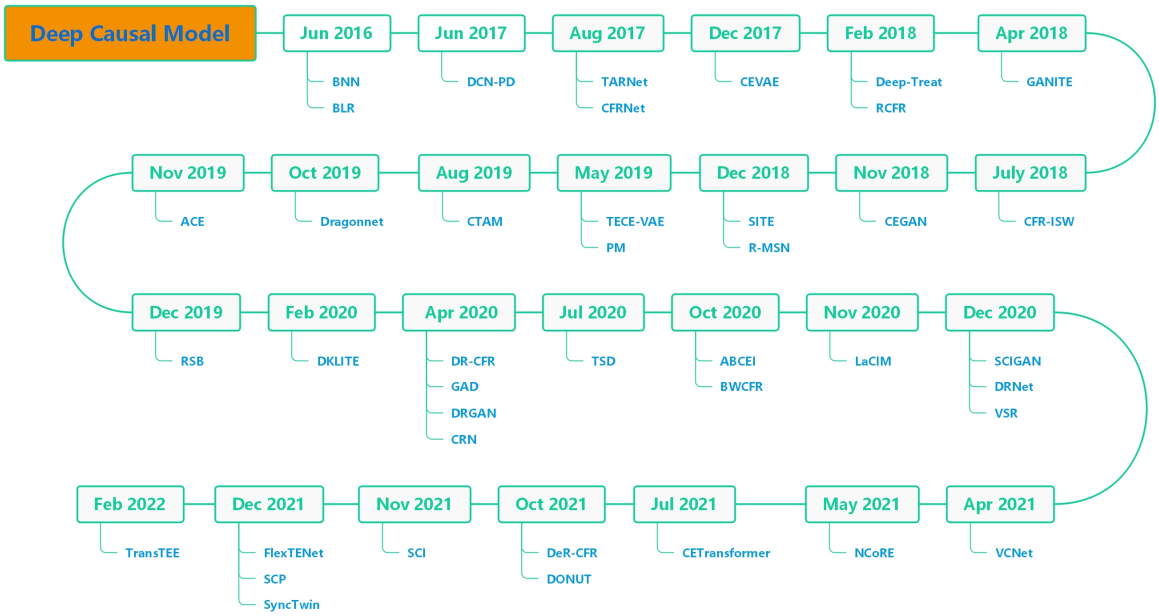

Figure 1: The Development of Deep Causal Models

In 2018, and going forward into 2019, there is an increasing interest in causal representation learning. In the beginning, the Deep-Treat[19] and RCFR[111] models are jointly proposed. After the launch of the GANITE[45] model, the use of generative adversarial model[112] architecture for counterfactual estimation becomes mainstream in the field of causal inference. In accordance with the previous work, CFR-ISW[113], CEGAN[72], SITE[44] are optimized. The R-MSN[74] model, implemented in December 2018, uses recurrent neural networks[114] to solve the problem of continuous dose of multi-treatment time series, which opened up the deep causal model. To tackle this problem, PM[42] and TECE[104] are proposed in May 2019 for causal effect estimation related to multiple discrete treatments. As a follow-up, the CTAM[34] begins to focus on estimating causal effects for textual data; the Dragonnet[71] introduces regularizations and propensity score networks into causal models for the first time; the ACE[55] attempts to extract fine-grained similarity information from representation space. For RSB's[115] December 2019 version, deep representation learning networks and PCC[116] regularization are used to decompose covariates, instrumental variables are used to control selection bias, and confounding and moderating factors are used for prediction.

Deep causal models are booming in 2020. Firstly, A DKLITE[56] model incorporates a deep kernel model and posterior variance regularization. Then, DR-CFR[117] applies three representation networks, two regression networks, and one prediction network, to decouple selection bias for covariates; GAD[118] then focuses on the causal effect of continuous dose treatment; DRGAN[119] defines an innovative generative adversarial network for fitting sample dose effect curves; and CRN[120] estimates time-varying treatment effects by combining counterfactual recurrent neural networks. After estimating time series causal effects under multi-cause confounding, TSD[121] turns to estimation of time series causal effects. In the latent representation space, ABCEI[122] balances the covariate distribution of treatment and control groups using GAN. Based on previous research, BWCFR[123], LaCIM[124] then optimize the structure idea. Furthermore, SCIGAN[73], DRNet[57] extend continuous dose to any number of treatment problems in 2020, and VSR[125] aggregates deep neural network latent variables in a reweighted manner.

From 2021 to 2022, causal models have become more innovative, open, and flexible. The VCNet[58] model, implements an estimator of continuous mean dose-response curves. As of May 2021, NCoRE[126] uses cross-treatment interaction modeling to understand the underlying causal processes that produce multiple treatment combinations.After that, CETransformer[127] uses Transformer[128] to characterize covariates, and the attention mechanism is focused on the correlation among covariates. Following that, DONUT[129] and DeR-CFR[70] optimize based on previous work. SCI[75] uses subspace theory for causal representation learning, broadening researchers' ideas. A multi-task adaptive learning architecture is proposed by FlexTENet[130].

Additionly, SCP[131] estimates multifactorial treatment effects using a two-step procedure. To construct this synthetic twin matching representation, SyncTwin[132] utilizes the temporal structure in the results. In the end, TransTEE[76] extends the representation distribution balance approach to continuous, structured, and dose-dependent treatments, making it more open-ended as a causal effect estimation problem.

The next section analyzes all models of the same category and makes a comparison based on the use of deep learning structures and the common ideas utilized by the models.

4.2 Classification

A brief overview of deep causal models over the past six years has been provided in the previous subsection. According to the type of the method, this subsection evaluates the relevant deep causal models. There are five categories of deep causal models currently available. Using the release time as the main line, we briefly describe the advantages of various algorithms. Figure 2 shows the detailed classification of each model.

A representation distribution balance is proposed in a method model in 2016. The hotbed of researchers' research for a long time has been this kind of method. The BNN[41] program's opening, in June 2016, lays the groundwork for such methods by relating causal inference to neighborhood adaptation. Afterwards, DCN-PD[110]introduces a deep multi-task neural network that performs counterfactual reasoning. As of August 2017, CFRNet[43] added a distance integral probability metric and an imbalance penalty based on BNN. In order to learn the optimal treatment strategy, Deep-Treat[19] uses an unbiased autoencoder network. Then, RCFR[111] and CFR-ISW[113] employ a re-weighting strategy to balance the spatial representation. From December 2018, SITE[44] retains the distribution of local similarity and data balance representing the spatial treatment and control groups. Beginning in November 2019, ACE[55] start focusing on finer grained similarity information in feature space. Following this, DKLite[56] learns to represent information about spatial domain overlap. By October 2020, BWCFR[123] spatially reweights domain-overlapping representations. Lastly, SCI[75] integrates the concept of subspace in November 2021 to create multi-space information supplements.

Covariate confounding learning is first proposed and applied by CEVAE[54] in December 2017. An objective of CEVAE is to mine associations of potential confounders and assess their impact on causal effects based on VAE. As part of its non-parametric estimation theory, Dragonnet[71] incorporates the regularization objective function and Propensity score prediction network in October 2019. Next, RSB[115] applies autoencoders and PCC regularization for covariate decomposition. With the implementation of covariate decoupling in April 2020, DR-CFR[117] implemented three representation networks, two regression networks, and two prediction networks. Afterwards, LaCIM[124] proposed two different versions of its latent causal model. Once this is done, VSR[125] uses a reweighted approach to isolate the promiscuous individuals. Optimizing the De-CFR algorithm by October 2021 is a major success for DeR-CFR[70]. During the same period, DONUT[129] estimates the mean treatment effect with a deep orthogonal network. Furthermore, FlexTENet[130] proposes a method for learning shared information across multiple tasks that adaptively identifies outcomes.

For the first time in April 2018, GANITE[45] applies Generative Adversarial Networks for the estimation of counterfactuals. Following this, CEGAN[72] uses generative adversarial networks to depict the spatial distribution by combining representation distribution balance with generative adversarial models. By October 2020, ABCEI[122] balances latent

representation spatial covariate distribution by using GAN networks. Besides, CETransformer[127] combines attention mechanisms with GAN networks in July 2021 to learn a balanced covariate representation.

As of December 2018, the R-MSN[74] model focuses on counterfactual recurrent networks under time series. Additionally, text sequences are subjected to a condition-based treatment-adversarial learning matching model proposed by CTAM[34] in August 2019. To estimate the long-term treatment effect, CRN[120] combined GANs with counterfactual recurrent neural networks in April 2020. After that, TSD[121] constructs an output RNN factor model based on multiple tasks. Last but not least, SyncTwin[132] builds a synthetic twin sample structure that enables counterfactual analysis of target patients.

Using matching ideas, PM[42] attempted to resolve the problem of discrete multi-discrete treatment in May 2019. At the same time, TECE-VAE[104] employs a variational autoencoder to extend the task embedding model to arbitrary subsets of multi-treatment situations. By April 2020, GAD[118] combines generative adversarial deconfounding algorithms for continuous treatment problems, removing associations between covariates and treatment variables. To create a complete dose-effect curve for each sample, DRGAN[119] utilizes the structure of generator, discriminator, and prediction network. With the addition of hierarchical discriminators in December 2020, SCIGAN[73] completes its original foundation. As well, DRNet[57] permits the drawing of individual dose-response curves for any number of treatments under continuous dose parameters. In April 2021, VCNet[58] proposed continuous prediction of head structure, emphasizing continuity of treatment. Following that, NCoRE[126] models cross-treatment interactions to determine the underlying causal generative processes driving multiple treatment combinations. With a two-step approach, SCP[131] estimates polycausal treatment effect in December 2021. TransTEE[76], the latest model, incorporates an attention mechanism in which balanced covariate representations are learned over GAN networks with the aim of treating discrete continuous or dose-related treatment problems.

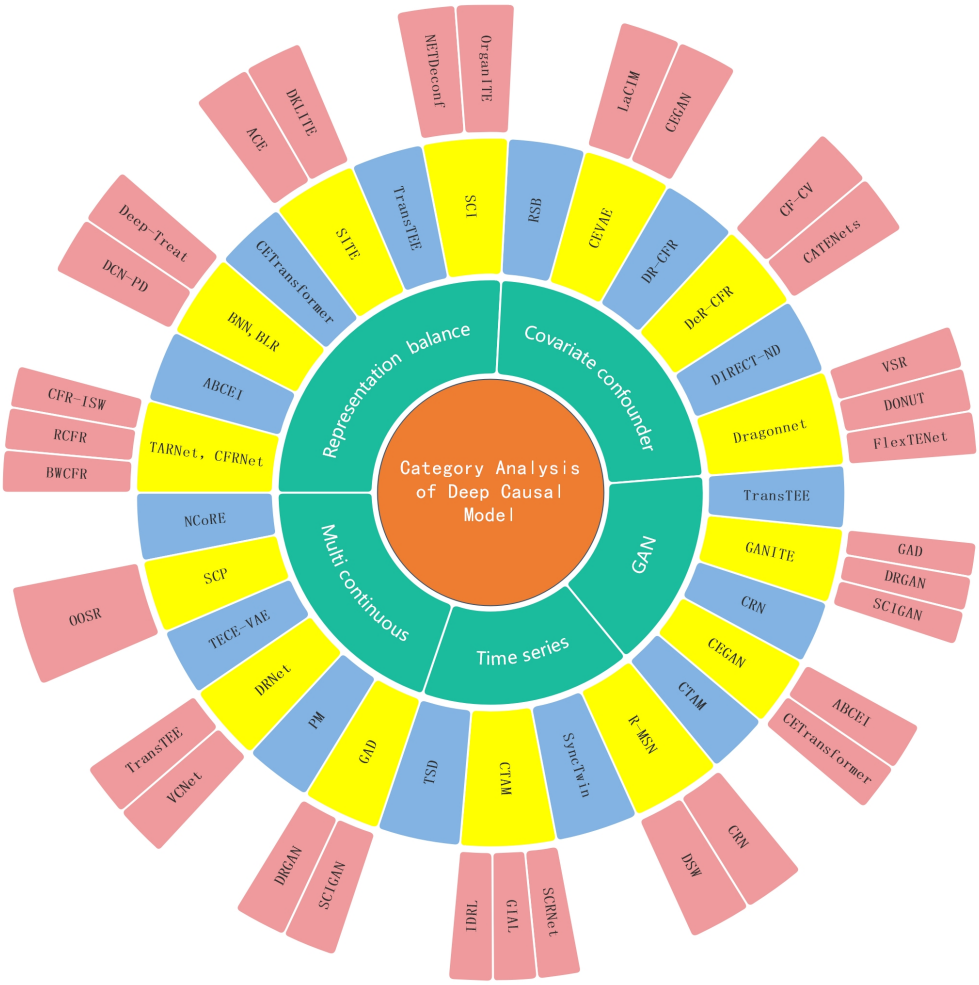

Figure 2: The deep causal model: A category analysis

As different application scenarios arise, the deep causal models use different iterations according to different strategies and methodologies. Please visit the next section for a detailed description and introduction to each deep causal model.

5 Methods

As more and more data accumulate in the fields of healthcare, education, economy, etc., deep learning approaches are increasingly used to infer causal relationships from counterfactual data. As opposed to existing deep causal models, which typically map covariates to a representation space, unbiased estimation of counterfactual data can be achieved using objective optimization functions. Current deep causal models mainly use these five optimization methods: 1) Representation of distribution balance methods; 2) Covariates confounding learning methods; 3) Methods based on Generative Adversarial Networks; 4) Time series causal estimation problem; 5) Methods based on multi-treatment and continuous-dose models. This section discusses in detail the current common methods for deep learning-based causal effects estimation, as well as the issues and challenges that these methods face.

5.1 Representation of distribution balance methods

Most statistical learning theories posit that test data and training data have independent and identical distributions, but in reality, the distributions of test data and training data are often related, but not identical. Solving this problem requires a machine learning model that learns causality rather than correlation in the field of causal inference. There is no standard treatment assignment strategy for observational data, unlike RCTs. Fact and counterfactual distributions are often different because of selection bias caused by known and unknown covariates. Hence, causal inference needs to be transformed into a domain adaptation problem to predict counterfactual outcomes by learning from factual data.

For counterfactual results to be predicted, effective feature representations are necessary, especially balanced distributions. According to Johansson et al, BNN[41] is an algorithmic framework for counterfactual reasoning that transforms the causal inference problem into a domain in Figure 3.

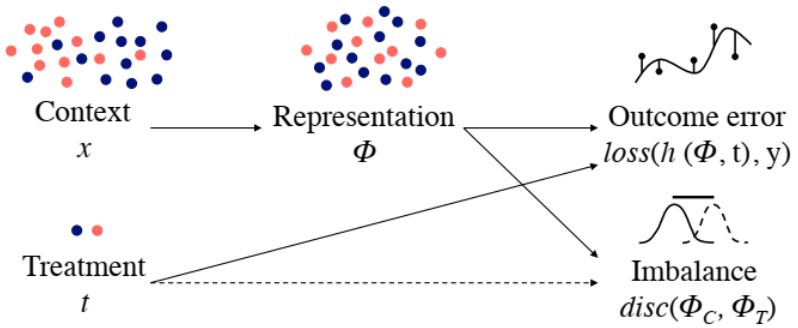

Figure 3: Representation Distribution Balance-Based Counterfactual Reasoning[41]

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. Here is the regression function:

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

An encoder network is represented by Φ, a predictor network by h, and a metric function is represented by $\mathrm{disc}_{\mathcal{H}}$ that represents the distance between the two distributions. In addition to representing the distribution distance in space, this function 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[43] proposes a CFRNet network structure based on BNN[41] algorithms and adopted MMD[133] and WASS[134, 135] 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 layers between each specific results layer, DCN-PD[110] combines multi-task deep neural networks with propensity score dropout.

Based on the CFRNet[43] model, RCFR[111] and CFR-ISW[113] use the Propensity score[136] to re-weight the representative spatial feature region and the sampling objective function; Atan et al. proposed an unbiased autoencoder network Deep-Treat[19] framework, which reduces the selection bias while reducing the loss of representation reconstruction, and applies a feedforward neural network to learn the optimal treatment strategy, weighing the selection bias and representation of the observation data as well as the information loss in space.

As a way to preserve local similarity and data balance representing the treatment group and the control group simultaneously, and to improve the individual treatment effect. Yao et al. proposes the SITE[44] method, which combines position-dependent depth metric PDDM with midpoint distance minimization MPDM into the representation space, and predictes potential results using a binary result network. In this case, the loss function is:

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

In the formula, $\mathcal{L}_{\mathrm{FL}}$ is the loss between predicted and observed factual outcomes, $\mathcal{L}_{\mathrm{PDDM}}$ and $\mathcal{L}_{\mathrm{PDDM}}$ are the loss functions of the PDDM and MPDM, respectively, and the last term is the L_2 regularization of the model parameter M.

According to SITE[44], ACE[55] proposes a balanced and adaptive similarity regularization structure to extract spatially fine-grained similarity information; DKLITE[56] proposes a deep kernel regression algorithm and a posterior regularization framework to learn the spatial domain overlap information; and BWCFR[123] re-weights the spatial feature distribution for the domain overlap region.

In many works, representation distribution balancing is combined with other domain ideas. An ABCEI[122] combines a GAN[112] with a mutual information estimator regularization structure to balance the covariate distributions of the treatment and control groups in the representation space; CETransformer[127] creates a balanced covariate representation using the attention mechanism; As TransTEE[76] extends the representation distribution balance method to Continuous, Structured, and Dose-Related treatments, it makes causal effect estimation a more open-ended problem; SCI[75] introduces 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 balanced representation and two specific representations. Afterwards, the public 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[137] provides network information that can be used to infer hidden confounders from observational data. A personalized treatment effect model that allocates treatments according to scarcity and estimates potential outcomes is proposed by OrganITE[138].

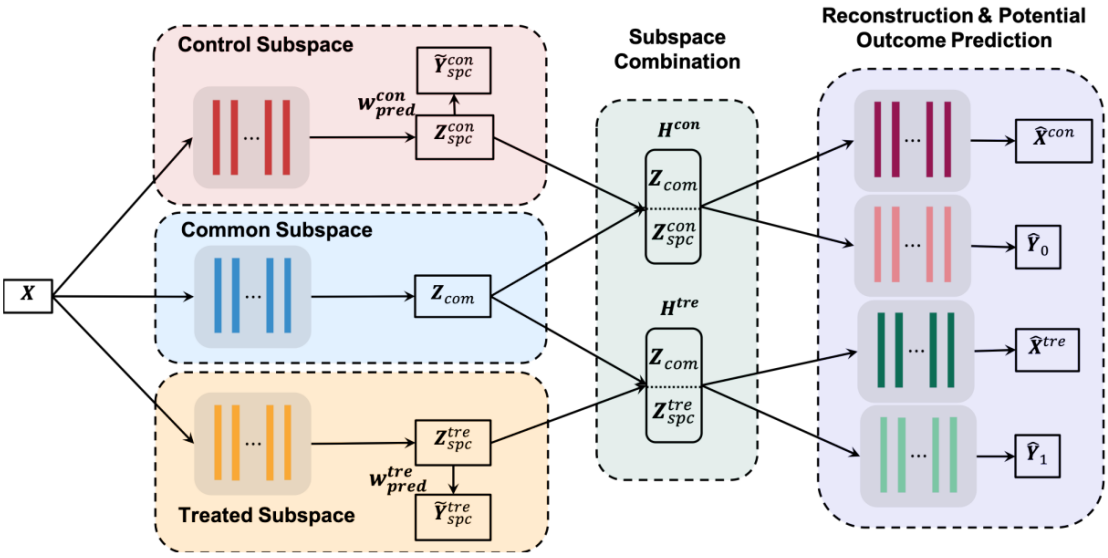

Figure 4: Structure of the SCI network[75]

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

5.2 Covariates confounding learning methods

The main issue in causal inference 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 covariate associations is an important method to learn counterfactual representations from observational data.

A CEVAE[54] model structure is first proposed by Louizos et al. to capture hidden confounding with VAEs [139, 140] in the presence of noise and uncertain confounding, to construct unobserved covariates and confounders, and to perform treatment and prediction. The causal correlation diagram is shown in Figure 5. Graphs can be represented as follows: t can represent drug treatment, y can represent mortality, z can represent socioeconomic status, and x can represent income and place of residence in the past year.

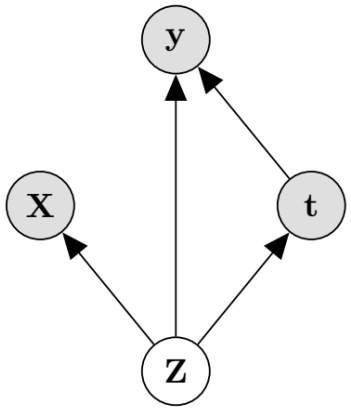

Figure 5: Causality diagram for CEVAE[54]

On the basis of TARNet[43]'s causal relationship diagram structure, do calculus is derived on 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 causal variational autoencoder has the following optimization function:

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

In the training set, input, treatment, and outcome random variables are observed at points $\mathbf{x}_i^*$, t_i^*, and y_i^*.

In response to CEVAE[54], Sun et al. proposes the LaCIM[124] latent causal model to avoid false associations and improve generalization ability of the model; CEGAN[72] uses GAN networks to identify potential confounders unbiasedly.

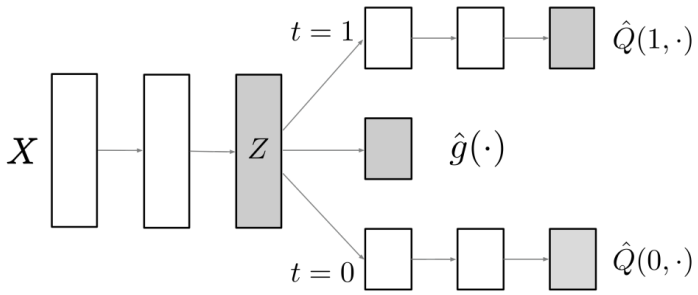

Figure 6: Structure of the Dragonnet network[71]

First-ever Dragonnet[71] proposed by Shi et al. adds regularization objective functions to nonparametric estimation theory and Propensity score prediction networks to CFRNet[43], thus making sure that the covariates are adjusted for treatment-related information in them. As a reference, Figure 6 shows the network structure. In accordance with Dragonnet[71], VSR[125] proposes a reweighting model that removes association processing and confounding factors, and uses a deep neural network to aggregate the density ratios of latent variables across the full variational distribution, which calculates the sample weight distribution; As part of the estimation process, DONUT[129] adds orthogonal constraints to the non-confounding factors in the loss function; An end-to-end regularization and reparameterization method called FlexTENet[130] learns a new architecture using multi-tasking to adaptively learn shared functions between causal structures. In DIRECT-ND[141], entanglement representation is solved through hybrid learning, and the multivariate causal effect estimation problem is studied from a new perspective. Additionally, VAE and GAN networks are added to the model to realize hybrid representation space learning.

In their first publication, Zhang et al. [115] proposes the RSB algorithm using autoencoder networks, PCC regularization, instrumental variables for balancing selection bias, and confounding variables and moderators for prediction. In Figure 7, DR-CFR[117] and DeR-CFR[70] are based on CFRNet[43], which uses three representation networks, two regression networks, and two prediction networks while removing covariate correlations.

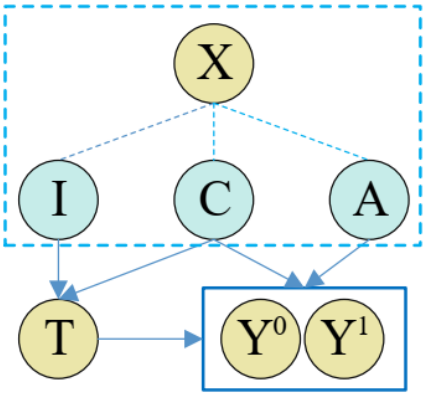

Figure 7: Causal framework for decoupling variables[70]

There are three possible factors contributing to the observed covariates X in the figure. The instrumental factor I, which only affects the treatment T; the confounding factor C, which causes the outcome Y along with the treatment T; and the adjustment factor A, which determines the outcome Y. Learning decomposition representation for counterfactual reasoning consists of the following steps[70]:

- Three decomposed representation networks for learning latent factors, one for each underlying factor: $I(X)$, $C(X)$, and $A(X)$.
- Three regularizers for confounder identification and balancing are presented: the first is to decompose A from X by considering $A(X) \perp\!\!\!\perp T$ and $A(X)$ should predict Y as accurately as possible; the second is to decompose I from X by constraining $I(X) \perp\!\!\!\perp Y \mid T$, and $I(X)$ should be predictive of T, based on Assumption 2; the last is designed for simultaneously balancing confounder $C(X)$ in different treatment arms.
- Two regression networks for potential outcome prediction, one for each treatment arm: $Y^0(C(X), A(X))$ and $Y^1(C(X), A(X))$.

Following is the orthogonal regularizer function used in the decomposition process:

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

In order to prevent the representation network from rejecting any input, the total of $\bar{W}_I$, $\bar{W}_C$ and $\bar{W}_A$ is constraint to be one. For a hard decomposition, the orthogonal regularizer ensures each variable in X can only flow into one representation network. Based on DeR-CFR, CATE's prediction performance is assessed using CF-CV[142], which selects the best model or hyperparameters from potential candidates. A meta-learning approach is combined with deep learning networks, theoretical reasoning, and optimal counterfactual information to guide principled algorithm design in paper[143].

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 method has some limitations in practical applications, as it requires decomposing covariates into reasonable explanations.

5.3 Methods based on Generative Adversarial Networks

In deep learning generative models, generative adversarial networks (GANs) can capture the uncertainty of counterfactual distributions using GANs[112]. The generator produces counterfactual results or a balanced distribution of the space, while the discriminator fits the non-uniformity of the treatment effects. In representation space, bias estimation or consistency in the distribution of control and treatment groups. As well as using factual data, GAN networks also consider the accuracy of counterfactual results when making causal inferences. 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[45] network to generate counterfactual results based on factual data and pass them to the ITE generator. As shown in Figure 8, the framework is structured as follows.

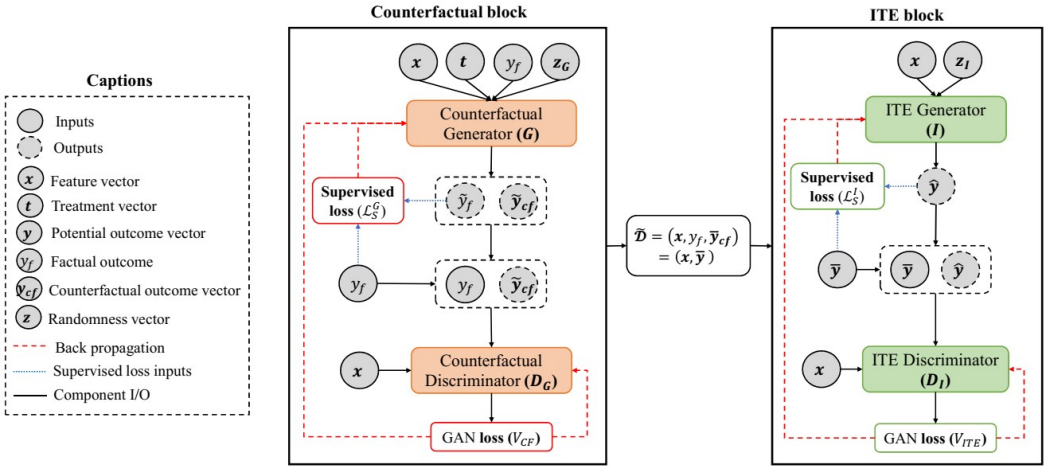

Figure 8: Structure of the GANITE frame[45]

From a given feature vector x, GANITE generates 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[72] applies the GAN network to balance the distribution between the spatial treatment group and the control group by learning 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 optimization function is used:

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

An encoder-decoder's joint distribution is represented by $q_E(\mathbf{z},\mathbf{x},t,\mathbf{y})$ and $q_p(\mathbf{z},\mathbf{x},t,\mathbf{y})$, while their probability estimates are represented by $(D(\hat{\mathbf{z}},\mathbf{x},t,\mathbf{y}))$ and $(1 - D(\mathbf{z},\mathbf{x},t,\hat{\mathbf{y}}))$, respectively. The discriminator determines which distribution the samples belong to.

In addition to GANITE and CEGAN, many works use GAN networks to estimate causal effects in other fields. As part of a generative adversarial framework, GAD[118] applies 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 treatments as well as consecutive doses treatment problems, DRGAN[119] proposes a model architecture consisting of a contrafactual generator, discriminator, and inference block; As a means of better coping with continuous intervention problems, SCIGAN[73] adds a hierarchical discriminator based on DRGAN; CTAM[34] applies generative

adversarial ideas to the treatment effect estimation of text sequence information, filters out information related to approximate instrumental variables when learning representations, and matches between the learned representations; To eliminate the association between treatment and patient history, CRN[120] creates a counterfactual recurrent neural network to reflect the time-varying treatment effect; In ABCEI[122], covariate distributions between the control and treatment groups are balanced with GAN networks, and a regularization function of mutual information estimators is added to reduce bias; To learn balanced covariate representations, CETransformer[127] combines Transformer[128] with attention mechanisms; With TransTEE[76], the covariate representation uses Transformer, the treatment effectiveness is estimated by the Propensity Score Network, and selection bias is overcome by the GAN network. The model can also be used for discrete, continuous, structured or dose-related treatments.

It is easy to extend the problem of individual treatment effect estimation to multiple interventions and continuous dose interventions using the GAN network, and it has a good effect on the balance of representation distribution and the generation of potential results. Other ideas combine to develop the GAN network into the latest concept.

5.4 Time series causal estimation problem

In treatment effect estimation, most models focus on numerical variables, and it is still unclear how to deal with textual information and time series information[144]. 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[114] have usually been combined to create counterfactual recurrent networks based on historical information.

The R-MSN[74] 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 9 illustrates the model's frame structure, which uses a recurrent edge network to remove time-dependent confounding, and a standard RNN structure to encode and decode.

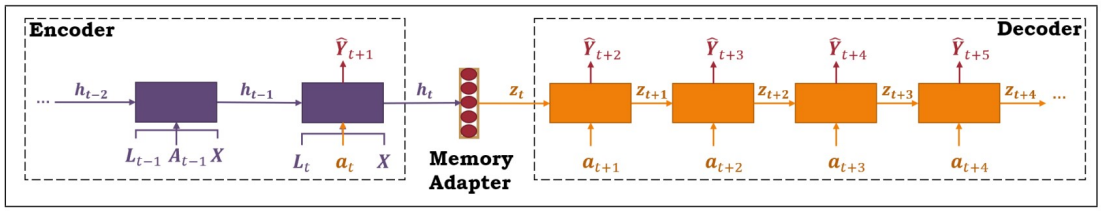

Figure 9: Framework for R-MSN[74]

To predict the causal effect, R-MSN uses the standard LSTM[145] structure, dividing multi-treatment and continuous intervention problems according to the corresponding time interval.

As a counterfactual recurrent network, CRN[120] constructs a treatment-invariant representation for each time step based on R-MSN[74], eliminating the patient's medical history association between treatment allocation and treatment allocation and balances time-varying confounding biases; In DSW[146], hidden confounders are inferred using recurrent weighted neural networks, reweighted using time-varying inverse probabilities, combined with current treatment assignment and historical information; In addition to building a multi-task output RNN factor model, TSD[121] allocates multiple treatments over time, estimates treatment effects with multi-cause hidden confounds, infers latent variables free of treatment, substitutes unobserved confounders with latent variables, and infers logistic regression in the absence of treatment; With SyncTwin[132], 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. proposes a matching treatment-adversarial learning CTAM[34] method that takes into account text sequence information. The CTAM filtering out approximate instrumental variables when learning representations, and matches between the learned representations to estimate treatment are shown in Figure 10.

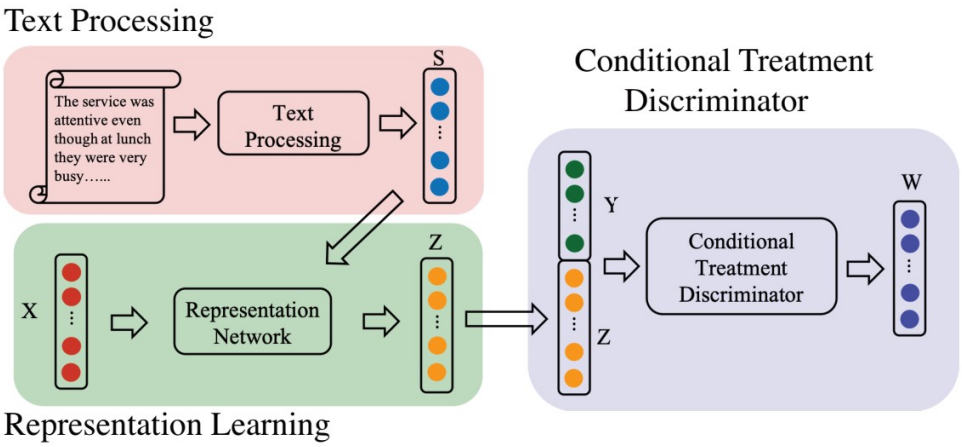

Figure 10: The structure of the CTAM model[34]

There are three main components of CTAM[34]: text processing, representation learning, and conditional treatment discrimination. In the first step, the text processing part transforms the original text into a vector representation S, concatenates S with non-text covariates X, and constructs a unified feature vector that transforms the input into the latent representation Z. As a next step, both Z and Y are fed into the conditioned treatment discriminator, and during the training process, a max-minimum game is played between the representation learning network and the conditioned treatment discriminator. In order to filter out information related to near instrumental variables, the representation learning network prevents the discriminator from assigning the appropriate treatment. As a last step, match the representation space Z.

Using the mutual information between global feature representations and individual feature representations as well as between feature representations and treatment assignment predictions, IDRL[147] proposes to learn Infomax and domain-independent representations. To maximize the capture of common prediction information among treatment and control groups, the influence of instrumental variables and irrelevant variables was filtered out. The SCRNet[148] divides covariates and estimates ITE with different types of variables. By identifying the imbalances in the network structure, the GIAL[149] model obtains more information from the network structure.

In the estimation of causal effects for text time series, it is often combined with problems related to multi-treatment and continuous-dose treatments. 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 job evaluation methods used in the industry.

5.5 Methods based on multi-treatment and continuous-dose models

Casual estimation for individual treatments focuses on solving binary treatment problems, and extending it to multiple treatments is computationally expensive. However, multiple treatments as well as prolonged usage of vasopressors have many applications, such as radiotherapy, chemotherapy and surgery for cancer treatment, as well as the use of continuous amounts of vasopressors[150] for many years. It is therefore beneficial to estimate the effects of ongoing interventions in these various treatment settings in order to make good long-term process decisions.

For the first time, Schwab et al. extend individual treatment estimation to multi-discrete treatment problems with the PM[42] algorithm. Counterfactual reasoning is utilized by PM in small batches by matching nearest neighbors samples. Despite the fact that it can be easily implemented and is compatible with a wide range of architectures, there is no need to increase computation complexity or other hyperparameters for treating any number of patients. By capturing Higher-order effects, TECE-VAE[104] models 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[118] can be used for continuous treatment problems, removing the association between covariates and treatment variables: A) Produce an unbiased distribution with no correlation between the covariates; B) Learn the sample weights, transfer observed data to the unbiased distribution, then de-obfuscate the data with generative adversarial networks.

An improved GAN model is proposed in DRGAN[119], which takes the form of a generator, discriminator, and prediction network to generate a complete dose-response curve for each sample, in which multi-treatment and continuous-dose treatment options are considered; By using a hierarchical discriminator based on DRGAN, SCIGAN[73] improves the model's ability to handle continuous intervention problems.

A set of open model benchmark parameters, including MISE, DPE, PE, and model selection criteria, are developed by DRNet[57], which allows the generation of dose-response curves for an unlimited number of treatments under continuous dose parameters. For VCNet[58], which utilizes a variable coefficient neural network, a continuous ADRF[151, 152, 153] estimator is automatically calculated for the continuous activation function, which prevents processing information from being lost. Moreover, the existing target regularization method is extended to obtain a double robust ADRF curve estimator. DRNet and VCNet model structure comparisons are shown in Figure 11.

As part of DRNet[57], 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[58] by paying closer attention to treatment continuity, and optimizing the individual prediction head into a mapping function of covariates that change with treatment.

In addition to SCIGAN[73] and VCNet[58], TransTEE[76] incorporates Transformer attention mechanism into it and expands it so that the model deals with discrete, continuous, and dose-related treatments.

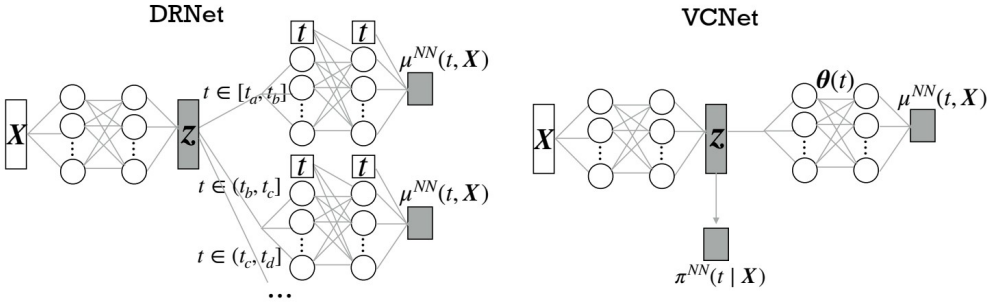

Figure 11: Network structure comparison between DRNet and VCNet[58]

As the first study of the multi-treatment combination problem, NCoRE[126] uses cross-treatment interaction modeling to infer causal generative processes underlying multiple treatment combinations, combining counterfactual representations learned in a treatment setting.

To estimate the multi-cause perturbation treatment effect in two steps. Prichard and colleagues proposes the idea of SCP[131] for the first time. To overcome confounding bias, the first step used a single-cause CATE estimator 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. According to the device, there is a theoretical basis for the proof, which is as follows:

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

In the first step of augmenting the dataset, outcomes and observations $Y\left(a_k\right)$ and $\mathbf{A}_{-k}\left(a_k\right)$ are added. 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 the formula as well as the multifactorial intervention on the left side of the formula effect in a way that enhances the generalizability of the estimator. In contrast to SCP, OOSR[154] proposes an outcome-oriented reweighting algorithm and a prediction model that emphasizes 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 applications in this area that need to be developed. It is still an urgent problem to solve how to formulate a unified causal effect measurement standard.

6 Guideline For Experiment

After understanding the deep causal model method, this section discusses the applicable experimental information, including examples of commonly used datasets and source codes for related experiments.

6.1 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. According to survey[2], we expand and classify relevant datasets, As shown in Figure 12, each commonly used dataset falls into the appropriate category. It is marked in yellow if a classic dataset is widely used. Table 2 summarizes the link to download the dataset and the classical methods related to it. Below are detailed descriptions of the general datasets for various methods.

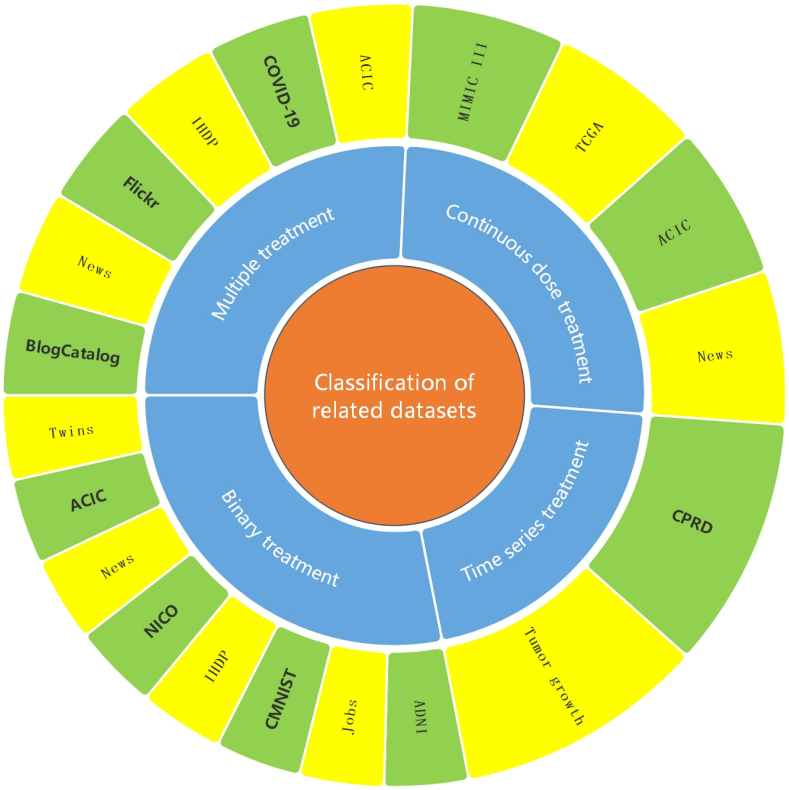

Figure 12: An analysis of available datasets by classification

IHDP. An infant health and development program[102] conducts a randomized controlled experiment that targets preterm infants with low birth weight to generate this dataset. Various aspects of the children and their mothers is measured as pre-treatment covariates, such as birth weight, head circumference, neonatal health index, prenatal care, mother's age, education, drugs, and alcohol. Intensive high-quality childcare is provided to infants in the treatment group and they also received specialist home visits[155]. The outcome is a score on the cognitive test for the infant. Moreover, to model selection bias, a biased subset of treatment groups needs to be removed.

Jobs. The employment data studied by Jobs in LaLonde (1986)[156] consisted of randomized data based on state-supported work programs and nonrandomized data from observational studies. The pre-treatment covariates are eight variables such as age, education, race, and income in 1974 and 1975. Treatment participants take part in vocational training, while control participants are not. The outcome is employment status.

Twins. The Twins dataset benchmark comes from data on twin births in the United States from 1989-1991[157]. An evaluation of 40 covariates pertaining to pregnancy, twin births, and parents was carried out for every pair of twins, including gestational weeks just before birth, quality of care during pregnancy, pregnancy risk factors (anemia, alcohol,

smoking, etc.), nursing, residence, and more. The outcome is a one-year mortality rate. It is estimated that the twins died within a year of being treated more heavily. A twin dataset is available with results from the treatment (heavier of the twins) and control groups (lighter of the twins). Selection bias is typically simulated by assigning treatments based on user-defined criteria.

News. The News dataset consists of 5000 randomly sampled news articles from the New York Times corpus. The news dataset contains data on media consumers' perceptions of news items. A sample is a news item consisting of word counts, the results are readers' opinions, and the available treatments are a variety of devices that can be used to view the news item, such as smartphones, tablets, computers, and TVs.

ACIC. A causal inference data analysis challenge has been held every year at the Atlantic Causal Inference Conference since 2016, which presents different data sets for a variety of causal inference problems. Here is a description of ACIC 2016 and ACIC 2018 datasets that are used in this article[71, 130]. A summary of the latest conference dataset can be found in the paper[158].

The ACIC 2016 consists of 77 datasets with different degrees of nonlinearity, sparsity, correlation between treatment assignment and outcome, and overlap between treatment effects. Covariates are derived from real data from the IHDP[102] dataset, which consists of 58 variables and 4802 samples[159]. The simulation model generates treatment, factual and counterfactual outcomes, while the selection bias is created by removing treated children who are mothers of nonwhites. The ACIC 2018 is a benchmarking framework for causal inference that is commonly used[160]. This is a collection of semi-synthetic datasets from related birth and infant death data[161]. It contains 63 datasets, each drawn randomly from a different distribution, that are then generated by a generative simulation process.

TCGA. As the world's largest and most comprehensive genomic database, The Cancer Genome Atlas (TCGA)[107] contains billions of genomes. A total of 9658 individuals are included in the TCGA[107] dataset, the treatment regimens are drug treatment, chemotherapy, and surgery, and the outcome is the risk of developing cancer after treatment.

PK-PD model of tumor growth. A model of pharmacokinetic-pharmacodynamics (PK-PD)[162] can be used to explore dose-response relationships and suggest optimal treatments[74]. Among its key characteristics are the combination of chemotherapy and radiotherapy effects, post-treatment cellular regeneration, patient death or recovery, and cancer-based different gaze distributions of tumor size at the diagnostic stage, which make this model an excellent model for treating non-small cell lung cancer patients. The PKPD model enables clinicians to explore hypotheses about dose-response relationships and suggest optimal treatment options[163, 164]. In the most classic example of PK-PD, tumor growth[162] can be predicted with time-dependent confounding by observing the expected response to treatment, chemotherapy, and radiotherapy.

MIMIC III. Medical Information Mart for Intensive Care(MIMIC III)[108] is a database of electronic health records from ICU patients. The benchmark consists of 7413 samples with 25 covariates after filtering for missing values. As far as treatment options go, antibiotics, vasopressors, and mechanical ventilators are the most common options in the ICU to treat patients with sepsis. A number of laboratory tests and vital signs measured over time are used to assess how antibiotics, vasopressors, and mechanical ventilators affected the following patient covariates: white blood cells, blood pressure, and oxygen saturation. A comprehensive and detailed description of the clinical data can be found in paper[165].

NICO. There is a bias in sample selection when using the image dataset NICO with context for object classification[124]. Cat or dog classification in the "animals" dataset in NICO a benchmark for non-i.i.d[166]. The parameters include the time of sampling, whether to sample, the context, and the semantic shape of cats and dogs, as well as the "grass" and "snow" environment.

CMNIST. A dataset of handwriting recognition with confusion bias(CMNIST) is based on MNIST[167] and labels the digits 0 to 4 and 5-9 for two outputs, namely green and red. A therapeutic input derived from color-evoked stimulus-related intensity parameters. Among covariates are number and color painting times, and whether or not to paint.

ADNI Alzheimer's Disease Neuroimaging Initiative(ADNI)[168] dataset has three latent representation outputs Alzheimer's Disease, Mild Cognitive Impairment and Normal Control. The covariates are age and TAU[169], which determine whether Magnetic resonance imaging should be used as an input to therapy.

COVID-19. During the first peak of the pandemic, dataset COVID-19[170, 171] Hospitalization in England Surveillance System (CHESS) collected individual-level risk factors, treatments, and outcomes from 3090 ICU patients. There are a number of covariates, including factors such as age and multiple morbidity, as well as treatment parameters, such as ventilation and antiviral drugs. The outcome is the length of stay in the intensive care unit[172].

CPRD. Clinical Practice Research Datalink (CPRD) contains records from NHS general practice clinics in the United Kingdom, covering approximately 6.9 pecent of the country's population[173]. National mortality records and hospital event statistics indicate that CPRD is associated with secondary care admissions. Low-density lipoprotein is measured after CPRD is initiated, and treatment initiation is defined as the date of first prescription. As time covariates, the following LDL risk factors is measured before treatment initiation: high-density lipoprotein cholesterol, blood pressure, pulse, creatinine, triglycerides, and smoking status. HPS registry participants are selected from 125,784 individuals who meet the eligibility criteria. A total of 17,371 treatment groups and 24,557 control groups are divided into three equally sized subsets for training, validation, and testing.

BlogCatalog BlogCatalog is an online community where users post blogs. In the dataset, each instance is a blogger[174]. Each edge represents a social relationship between two bloggers. Blog descriptions contain keywords represented as a bag-of-words. Blog reader opinions as input, whether blog-created content gets more comments on mobile or desktop as therapy, research on the effect of getting more reader opinions on mobile (than desktop) on individual therapeutic effects of reader opinions. A blogger belongs to the treated group (control group) if people reads more on a mobile device than on a desktop device.

Flickr Flickr is an online social networking site where users can share photos and videos[175]. The dataset consists of instances representing users, and edges representing social relationships between them. Tags of interest are represented by the features of each user. General settings and assumptions are the same as for BlogCatalog dataset.

Table 2: Related Classic Methods and Links to Datasets

Dataset	Link	Method
IHDP	https://www.fredjo.com/files	[41, 110, 43, 54, 19, 111, 45, 44, 42, 104, 34, 71, 55, 115, 56, 117, 122, 123, 58, 127, 70, 130, 143, 127, 76, 148]
Jobs	http://users.nber.org/rdehejia/data/nswdata2.html	[43, 54, 45, 44, 55, 56, 122, 127, 129, 75, 148, 147]
Twins	www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html	[54, 45, 72, 44, 55, 56, 118, 122, 127, 129]
News	https://archive.ics.uci.edu/ml/datasets/bag+of+words	[41, 42, 104, 34, 119, 73, 75, 58, 76, 147]
ACIC	https://www.synapse.org/ACIC2018Challenge	[71, 130, 122, 176, 177, 178, 179, 180, 181, 182]
TCGA	https://gdc.cancer.gov/	[119, 73, 75, 76, 154, 183, 184, 185, 186, 187]
Tumor Growth	www.nature.com/scientificreports/	[74, 120, 188, 189, 190, 191]
MIMIC III	https://mimic.physionet.org/	[121, 122, 73, 57, 75, 192]
NICO	https://www.dropbox.com/sh/8mouawi5guaupyb/AAD4fdySrA6fn3PgSmhKwFgvadl=0	[124, 193, 194, 195, 196, 197]
CMNIST	1https://trends.google.com/trends/explore?date=all&q=mnist	[124, 198, 199, 200, 201, 202]
ADNI	www.loni.ucla.edu/ADNI	[124, 203, 204, 205, 206]
COVID-19	https://www.heywhale.com/mw/dataset/5e8ee81fe7ec38002d00f9cb	[131, 207, 208]
CPRD	https://academic.oup.com/ije/article/44/3/827/632531	[132, 209, 210, 211, 212, 213, 214]
BlogCatalog	https://www.blogcatalog.com/	[137, 215, 216]
Flickr	https://www.flickr.com	[137]

6.2 Codes

This subsection summarizes the relevant available datasets and codes according to method classification. Links to source code for specific methods are provided in the Table 3. The table also displays the classic data sets used by each deep network model method to help you understand the specific use of the papers.

Table 3: Available Codes and Datasets of Methods

Method	Datasets	Framework	Link
DCN-PD[110]	IHDP	Pytorch	https://github.com/Shantanu48114860/Deep-Counterfactual-Networks-with-Propensity-Dropout
BNN[41],CFRNet[43]	IHDP,Jobs,News	Tensorflow	https://github.com/oddrose/cfrnet
CEVAE[54]	IHDP,Twins,Jobs	Tensorflow	https://github.com/AMLab-Amsterdam/CEVAE
GANITE[45]	IHDP,Twins,Jobs	Tensorflow	https://github.com/jsyoon0823/GANITE
SITE[44]	IHDP,Twins,Jobs	Tensorflow	https://github.com/Osier-Yi/SITE
R-MSN[74]	PK-PD model of tumor growth	Tensorflow	https://github.com/sjblim/rmsn_nips_2018
PM[42]	IHDP,News	Tensorflow	https://github.com/d916b/perfect_match
Dragonnet[71]	IHDP,ACIC	Tensorflow	https://github.com/claudiashi57/dragonnet
DKLITE[56]	IHDP,Twins,Jobs	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/dklite
CRN[120]	PK-PD model of tumor growth	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/counterfactual_recurrent_network
TSD[121]	MIMIC III	Tensorflow	https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/time_series_deconfounder
ABCEI[122]	IHDP,Twins,Jobs,ACIC,MIMIC III	Tensorflow	https://github.com/octeufer/Adversarial-Balancing-based-representation-learning-for-Causal-Effect-Inference
LaCIM[124]	NICO,CMNIST,ADNI	Pytorch	https://github.com/wubotong/LaCIM
SCIGAN[73]	TCGA,News,MIMIC III	Tensorflow	https://github.com/ioanabica/SCIGAN
DRNet[57]	TCGA,News,MIMIC III	Tensorflow	https://github.com/d909b/drnet
VCNet[58]	IHDP,News	Pytorch	https://github.com/lushleaf/varying-coefficient-net-with-functional-tr
DeR-CFR[70]	IHDP	Tensorflow	https://github.com/anpwu/DeR-CFR
DONUT[129]	IHDP,Twins,Jobs	Tensorflow	https://github.com/tobhatt/donut
FlexTENet[130],CATENets[143]	IHDP,Twins,ACIC	Jax,Pytorch	https://github.com/AliciaCurth/CATENets
SCP[131]	COVID-19	Pytorch	https://github.com/vanderschaarlab/Single-Cause-Perturbation-NeurIPS-2021
SyncTwin[132]	CPRD	Pytorch	https://github.com/vanderschaarlab/SyncTwin-NeurIPS-2021
TransTEE[76]	IHDP,News,TCGA	Pytorch	https://github.com/hlzhang109/TransTEE
CF-CV[142]	IHDP	Tensorflow	https://github.com/usaito/counterfactual-cv

By combining related methods, data sets, and source code, we can more easily identify the innovation points in each model, realize the operation and modification of source code, and propose meaningful model strategy research. As an example, the covariate decomposition is applied to the Dragonnet[71] network model when combined with the DeR-CFR[70] model to optimize the algorithm. Fit continuous dose estimation curves more accurately by applying the TransTEE[76] attention mechanism to the representation balance part of VCNet[58] or DRNet[57]. Researchers can promote the rapid development of the field of causal inference through the combination and innovation of various algorithms.

7 Conclusions

Deep causal models have become increasingly popular as a research topic because of the development of causal inference. It is possible to improve causal effect estimation accuracy and unbiasedness by applying deep learning network models to causal inference. Additionally, deep learning networks can be optimized and improved by applying profound theories in causal reasoning. The paper presents the development of deep causal models and the evolution of various methods, beginning with relevant background information in the field of causal reasoning, as well as detailed descriptions of methodologies, including definitions, assumptions, sample measurement standards, etc. a category description and a thorough comparison and summation of the various approaches. Finally, the paper also lists available benchmark datasets and open source codes for these methods.

References

[1] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

[2] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[3] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. Causal inference via sparse additive models with application to online advertising. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[4] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 67–76, 2015.

[5] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*, pages 3768–3774, 2016.

[6] Christian Fong, Chad Hazlett, and Kosuke 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.

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

[8] Bowen Yuan, Jui-Yang Hsia, Meng-Yuan Yang, Hong Zhu, Chih-Yao Chang, Zhenhua Dong, and Chih-Jen Lin. Improving ad click prediction by considering non-displayed events. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 329–338, 2019.

[9] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[10] Akos Lada, Alexander Peysakhovich, Diego Aparicio, and Michael Bailey. Observational data for heterogeneous treatment effects with application to recommender systems. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 199–213, 2019.

[11] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

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

[13] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. A causal perspective to unbiased conversion rate estimation on data missing not at random. *arXiv preprint arXiv:1910.09337*, 2019.

[15] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 104–112, 2018.

[16] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pages 6638–6647. PMLR, 2019.

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

[18] Ronald C Kessler, Robert M Bossarte, Alex Luedtke, Alan M Zaslavsky, and Jose R Zubizarreta. Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy*, 120:103412, 2019.

[19] Onur Atan, James Jordon, and Mihaela Van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[20] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[21] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

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

[23] Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

[24] Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36. JMLR Workshop and Conference Proceedings, 2012.

[25] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

[26] Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

[27] Hao Zou, Kun Kuang, Boqi Chen, Peixuan Chen, and Peng Cui. Focused context balancing for robust offline policy evaluation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 696–704, 2019.

[28] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

[29] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 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)*, pages 1615–1625, 2018.

[30] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

[31] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*, 2014.

[32] Zach Wood-Doughty, Ilya Shpitser, and Mark 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*, volume 2018, page 4586. NIH Public Access, 2018.

[33] Katherine A Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*, 2020.

[34] Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. On the estimation of treatment effect with text covariates. In *International Joint Conference on Artificial Intelligence*, 2019.

[35] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[36] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019.

[37] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019.

[38] Vedika Agarwal, Rakshith Shetty, and Mario 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*, pages 9690–9698, 2020.

[39] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 378–379, 2020.

[40] Siyuan Zhao and Neil Heffernan. Estimating individual treatment effect from educational studies with residual counterfactual networks. *International Educational Data Mining Society*, 2017.

[41] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

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

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

[44] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[46] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1617–1626, 2018.

[47] Mehdi Gheisari, Guojun Wang, and Md Zakirul Alam 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)*, volume 2, pages 173–180. IEEE, 2017.

[48] Reinaldo Padilha França, Ana Carolina Borges Monteiro, Rangel Arthur, and Yuzo Iano. An overview of deep learning in big data, image, and signal processing in the modern digital age. *Trends in Deep Learning Methodologies*, pages 63–87, 2021.

[49] Qingchen Zhang, Laurence T Yang, Zhikui Chen, and Peng Li. A survey on deep learning for big data. *Information Fusion*, 42:146–157, 2018.

[50] Bilal Jan, Haleem Farman, Murad Khan, Muhammad Imran, Ihtesham Ul Islam, Awais Ahmad, Shaukat Ali, and Gwanggil Jeon. Deep learning in big data analytics: a comparative study. *Computers & Electrical Engineering*, 75:275–287, 2019.

[51] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. 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)*, pages 1–6. IEEE, 2017.

[52] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

[53] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[54] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[55] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1432–1437. IEEE, 2019.

[56] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.

[57] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

[58] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

[59] Naomi Altman and Martin Krzywinski. Points of significance: Association, correlation and causation. *Nature methods*, 12(10), 2015.

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

[61] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[62] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2021.

[63] Amar Bhide, Prakesh S Shah, and Ganesh Acharya. A simplified guide to randomized controlled trials. *Acta obstetricia et gynecologica Scandinavica*, 97(4):380–387, 2018.

[64] Christine M Steeger, Pamela R Buckley, Fred C Pampel, Charleen J Gust, and Karl G Hill. Common methodological problems in randomized controlled trials of preventive interventions. *Prevention Science*, 22(8):1159–1172, 2021.

[65] Stephen A Kutcher, James M Brophy, Hailey R Banack, Jay S Kaufman, and Michelle Samuel. Emulating a randomised controlled trial with observational data: an introduction to the target trial framework. *Canadian Journal of Cardiology*, 37(9):1365–1377, 2021.

[66] Gemma Hammerton and Marcus R Munafò. Causal inference with observational data: the need for triangulation of evidence. *Psychological medicine*, 51(4):563–578, 2021.

[67] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.

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

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

[70] Anpeng Wu, Kun Kuang, Junkun Yuan, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representation for counterfactual inference. *arXiv preprint arXiv:2006.07040*, 2020.

[71] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[72] Changhee Lee, Nicholas Mastronarde, and Mihaela van der Schaar. Estimation of individual treatment effect in latent confounder models via adversarial learning. *arXiv preprint arXiv:1811.08943*, 2018.

[73] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

[74] Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. *advances in neural information processing systems*, 31, 2018.

[75] Liuyi Yao, Yaliang Li, Sheng Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Sci: Subspace learning based counterfactual inference for individual treatment effect estimation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3583–3587, 2021.

[76] Yi-Fan Zhang, Hanlin Zhang, Zachary C Lipton, Li Erran Li, and Eric P Xing. Can transformers be strong treatment effect estimators? *arXiv preprint arXiv:2202.01336*, 2022.

[77] Paul Muentener and Elizabeth Bonawitz. The development of causal reasoning. 2018.

[78] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022.

[79] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020.

[80] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[81] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

[82] Hang Chen, Keqing Du, Xinyu Yang, and Chenguang Li. A review and roadmap of deep learning causal discovery in different variable paradigms, 2022.

[83] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions, 2022.

[84] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *arXiv preprint arXiv:1704.02801*, 2017.

[85] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

[86] Marc Höfler. Causal inference based on counterfactuals. *BMC medical research methodology*, 5(1):1–12, 2005.

[87] Markus Gangl. Causal inference in sociological research. *Annual review of sociology*, 36:21–47, 2010.

[88] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[89] Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? university of western ontario und university of pennsylvania. Technical report, mimeo, 2002.

[90] Alexis Hannart, J Pearl, FEL Otto, P Naveau, and M Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99–110, 2016.

[91] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[92] Pengyuan Wang, Wei Sun, Dawei Yin, Jian Yang, and Yi Chang. Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 67–76, 2015.

[93] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*, pages 3768–3774, 2016.

[94] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[95] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 265–274, 2017.

[96] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[97] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[98] Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *NIPS*, 2017.

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

[100] Donna Döpp-Zemel and AB Johan Groeneveld. High-dose norepinephrine treatment: determinants of mortality and futility in critically ill patients. *American Journal of Critical Care*, 22(1):22–32, 2013.

[101] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020.

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

[103] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[104] Shiv Kumar Saini, Sunny Dhamnani, Akil Arif Ibrahim, and Prithviraj Chavan. Multiple treatment effect estimation using deep generative model with task embedding. In *The World Wide Web Conference*, pages 1601–1611, 2019.

[105] Zhaozhi Qian, Ahmed M Alaa, Mihaela van der Schaar, and Ari Ercole. Between-centre differences for covid-19 icu mortality from early data in england. *Intensive care medicine*, 46(9):1779–1780, 2020.

[106] Nicolai Haase, Ronni Plovsing, Steffen Christensen, Lone Musaeus Poulsen, Anne Craveiro Brøchner, Bodil Steen Rasmussen, Marie Helleberg, Jens Ulrik Staehr Jensen, Lars Peter Kloster Andersen, Hanna Siegel, et al. 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.

[107] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[108] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[109] Niall D Ferguson, Eddy Fan, Luigi Camporota, Massimo Antonelli, Antonio Anzueto, Richard Beale, Laurent Brochard, Roy Brower, Andrés Esteban, Luciano Gattinoni, et al. The berlin definition of ards: an expanded rationale, justification, and supplementary material. *Intensive care medicine*, 38(10):1573–1582, 2012.

[110] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

[111] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

[112] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[113] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

[114] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

[115] Zichen Zhang, Qingfeng Lan, Lei Ding, Yue Wang, Negar Hassanpour, and Russell Greiner. Reducing selection bias in counterfactual reasoning for individual treatment effects estimation. *arXiv preprint arXiv:1912.09040*, 2019.

[116] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[117] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[118] Yunzhe Li, Kun Kuang, Bo Li, Peng Cui, Jianrong Tao, Hongxia Yang, and Fei Wu. Continuous treatment effect estimation via generative adversarial de-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 4–22. PMLR, 2020.

[119] Ioana Bica, James Jordon, and Mihaela van der Schaar. Individualised dose-response estimation using generative adversarial nets. *ICLR 2020 Conference Blind Submission*, 2019.

[120] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

[121] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pages 884–895. PMLR, 2020.

[122] Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

[123] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

[124] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-yan Liu. Latent causal invariant model. *arXiv preprint arXiv:2011.02203*, 2020.

[125] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33:19705–19715, 2020.

[126] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint arXiv:2103.11175*, 2021.

[127] Zhenyu Guo, Shuai Zheng, Zhizhe Liu, Kun Yan, and Zhenfeng Zhu. Cetransformer: Casual effect estimation via transformer based representation learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 524–535. Springer, 2021.

[128] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[129] Tobias Hatt and Stefan Feuerriegel. Estimating average treatment effects via orthogonal regularization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 680–689, 2021.

[130] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[131] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. *Advances in Neural Information Processing Systems*, 34, 2021.

[132] Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. Synctwin: Treatment effect estimation with longitudinal outcomes. *Advances in Neural Information Processing Systems*, 34, 2021.

[133] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[134] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[135] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[136] Paul R Rosenbaum. Propensity score. *Wiley Encyclopedia of Clinical Trials*, 2007.

[137] Ruocheng Guo, Jundong Li, and Huan Liu. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 232–240, 2020.

[138] Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. *Advances in neural information processing systems*, 33:20037–20050, 2020.

[139] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[140] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[141] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Multi-cause effect estimation with disentangled confounder representation. In *IJCAI*, pages 2790–2796, 2021.

[142] Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pages 8398–8407. PMLR, 2020.

[143] Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[144] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, pages 1–45, 2021.

[145] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[146] Ruoqi Liu, Changchang Yin, and Ping Zhang. Estimating individual treatment effects with time-varying confounders. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 382–391. IEEE, 2020.

[147] Zhixuan Chu, Stephen L Rathbun, and Sheng 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)*, pages 433–441. SIAM, 2022.

[148] Liu Qidong, Tian Feng, Ji Weihua, and Zheng Qinghua. A new representation learning method for individual treatment effect estimation: Split covariate representation network. In *Asian Conference on Machine Learning*, pages 811–822. PMLR, 2020.

[149] Zhixuan Chu, Stephen L Rathbun, and Sheng 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)*, pages 433–441. SIAM, 2022.

[150] Donna Döpp-Zemel and AB Johan Groeneveld. High-dose norepinephrine treatment: determinants of mortality and futility in critically ill patients. *American Journal of Critical Care*, 22(1):22–32, 2013.

[151] BN Prichard and PM Gillam. 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.

[152] ARTHUR B Schneider, ELAINE Ron, Jay Lubin, Marilyn Stovall, and Theresa C Gierlowski. 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.

[153] Timothy J Threlfall and Dallas R English. Sun exposure and pterygium of the eye: a dose-response curve. *American journal of ophthalmology*, 128(3):280–287, 1999.

[154] Hao Zou, Bo Li, Jiangang Han, Shuiping Chen, Xuetao Ding, and Peng Cui. Counterfactual prediction for outcome-oriented treatments. In *International Conference on Machine Learning*, pages 27693–27706. PMLR, 2022.

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

[156] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

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

[158] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[159] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

[160] Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

[161] Marian F MacDorman and Jonnae O Atkinson. Infant mortality statistics from the linked birth/infant death data set–1995 period data. 1998.

[162] Changran Geng, Harald Paganetti, and Clemens 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.

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

[164] Miro J Eigenmann, Nicolas Frances, Thierry Lavé, and Antje-Christine Walz. Pkpd modeling of acquired resistance to anti-cancer drug treatment. *Journal of pharmacokinetics and pharmacodynamics*, 44(6):617–630, 2017.

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

[166] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021.

[167] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[168] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. 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.

[169] Christian Humpel and Tanja Hochstrasser. Cerebrospinal fluid and blood biomarkers in alzheimer's disease. *World journal of psychiatry*, 1(1):8, 2011.

[170] Nicolai Haase, Ronni Plovsing, Steffen Christensen, Lone Musaeus Poulsen, Anne Craveiro Brøchner, Bodil Steen Rasmussen, Marie Helleberg, Jens Ulrik Staehr Jensen, Lars Peter Kloster Andersen, Hanna Siegel, et al. 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.

[171] Zhaozhi Qian, Ahmed M Alaa, Mihaela van der Schaar, and Ari Ercole. Between-centre differences for covid-19 icu mortality from early data in england. *Intensive Care Medicine*, 46(9):1779–1780, 2020.

[172] Chintan Ramani, Eric M Davis, John S Kim, J Javier Provencio, Kyle B Enfield, and Alex Kadl. Post-icu covid-19 outcomes: a case series. *Chest*, 159(1):215–218, 2021.

[173] Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd Van Staa, and Liam Smeeth. Data resource profile: clinical practice research datalink (cprd). *International journal of epidemiology*, 44(3):827–836, 2015.

[174] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. Does similarity breed connection?-an investigation in blogcatalog and last. fm communities. In *2010 IEEE Second International Conference on Social Computing*, pages 570–575. IEEE, 2010.

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

[176] Kwonsang Lee, Falco J Bargagli-Stoffi, and Francesca Dominici. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

[177] Falco J Bargagli-Stoffi, Kristof De-Witte, and Giorgio Gnecco. Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach. *arXiv preprint arXiv:1905.12707*, 2019.

[178] Kosuke Imai and Michael Lingzhi Li. Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, pages 1–15, 2021.

[179] Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. 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.

[180] David Jensen. Comment: Strengthening empirical evaluation of causal inference methods. *Statistical Science*, 34(1):77–81, 2019.

[181] Lu Cheng, Ruocheng Guo, Raha Moraffah, K Selçuk Candan, Adrienne Raglin, and Huan Liu. A practical data repository for causal learning with big data. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 234–248. Springer, 2019.

[182] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[183] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

[184] Jonathan D Young, Bryan Andrews, Gregory F Cooper, and Xinghua Lu. Learning latent causal structures with a redundant input neural network. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 62–91. PMLR, 2020.

[185] Junpeng Zhang, Vu Viet Hoang Pham, Lin Liu, Taosheng Xu, Buu Truong, Jiuyong Li, Nini Rao, and Thuc Duy Le. Identifying mirna synergism using multiple-intervention causal inference. *BMC bioinformatics*, 20(23):1–11, 2019.

[186] Christopher Schmidt, Johannes Huegle, Siegfried Horschig, and Matthias Uflacker. Out-of-core gpu-accelerated causal structure learning. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 89–104. Springer, 2019.

[187] Jonathan Young. *Deep Learning for Causal Structure Learning Applied to Cancer Pathway Discovery*. PhD thesis, University of Pittsburgh, 2020.

[188] Cheng-Ying Chou, Wan-I Chang, Tzyy-Leng Horng, and Win-Li Lin. Numerical modeling of nanodrug distribution in tumors with heterogeneous vasculature. *PLoS One*, 12(12):e0189802, 2017.

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

[190] Spyridon Patmanidis, Alexandros C Charalampidis, Ioannis Kordonis, Katerina Strati, Georgios D Mitsis, and George P Papavassilopoulos. Individualized growth prediction of mice skin tumors with maximum likelihood estimators. *Computer Methods and Programs in Biomedicine*, 185:105165, 2020.

[191] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. 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.

[192] Kaouter Karboub and Mohamed Tabaa. A machine learning based discharge prediction of cardiovascular diseases patients in intensive care units. In *Healthcare*, volume 10, page 966. Multidisciplinary Digital Publishing Institute, 2022.

[193] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

[194] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems*, 34:16846–16859, 2021.

[195] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

[196] Zhaoquan Yuan, Xiao Peng, Xiao Wu, Bing-kun Bao, and Changsheng Xu. Meta-learning causal feature selection for stable prediction. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[197] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. Meta-causal feature learning for out-of-distribution generalization. *arXiv preprint arXiv:2208.10156*, 2022.

[198] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019.

[199] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 33:5453–5467, 2020.

[200] Chao-Han Huck Yang, Yi-Chieh Liu, Pin-Yu Chen, Xiaoli Ma, and Yi-Chang James Tsai. When causal intervention meets adversarial examples and image masking for deep neural networks. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3811–3815. IEEE, 2019.

[201] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.

[202] Thien Q Tran, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Unsupervised causal binary concepts discovery with vae for black-box model explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9614–9622, 2022.

[203] Sebastian Pölsterl and Christian Wachinger. Estimation of causal effects in the presence of unobserved confounding in the alzheimer's continuum. In *International Conference on Information Processing in Medical Imaging*, pages 45–57. Springer, 2021.

[204] Baoliang Zhang, Xiaoxin Guo, Qifeng Lin, Haoren Wang, and Songbai Xu. Counterfactual inference graph network for disease prediction. *Knowledge-Based Systems*, page 109722, 2022.

[205] Christian Wachinger, Benjamin Gutierrez Becker, Anna Rieckmann, and Sebastian Pölsterl. Quantifying confounding bias in neuroimaging datasets with causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 484–492. Springer, 2019.

[206] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. Harmonization with flow-based causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 181–190. Springer, 2021.

[207] Nina Van Goethem, Ben Serrien, Mathil Vandromme, Chloé Wyndham-Thomas, Lucy Catteau, Ruben Brondeel, Sofieke Klamer, Marjan Meurisse, Lize Cuypers, Emmanuel André, et al. 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.

[208] Harrison Wilde, Thomas Mellan, Iwona Hawryluk, John M Dennis, Spiros Denaxas, Christina Pagel, Andrew Duncan, Samir Bhatt, Seth Flaxman, Bilal A Mateen, et al. 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.

[209] Emilia Gvozdenović, Lucio Malvisi, Elisa Cinconze, Stijn Vansteelandt, Phoebe Nakanwagi, Emmanuel Aris, and Dominique Rosillon. 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.

[210] Marie-Laure Charpignon, Bella Vakulenko-Lagun, Bang Zheng, Colin Magdamo, Bowen Su, Kyle Evans, Steve Rodriguez, Artem Sokolov, Sarah Boswell, Yi-Han Sheu, et al. Drug repurposing of metformin for alzheimer's disease: Combining causal inference in medical records data and systems pharmacology for biomarker identification. *medRxiv*, 2021.

[211] Shishir Rao, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Yikuan Li, Rema Ramakrishnan, Abdelaali Hassaine, Dexter Canoy, and Kazem Rahimi. Targeted-behrt: Deep learning for observational causal inference on longitudinal electronic health records. *arXiv preprint arXiv:2202.03487*, 2022.

[212] Jie Zhu and Blanca Gallego. Cds–causal inference with deep survival model and time-varying covariates. *arXiv preprint arXiv:2101.10643*, 2021.

[213] Janie Coulombe. *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.

[214] Antonia Marsden. *Causal Modelling in Stratified and Personalised Health: Developing Methodology for Analysis of Primary Care Databases in Stratified Medicine*. The University of Manchester (United Kingdom), 2016.

[215] Jundong Li, Ruocheng Guo, Chenghao Liu, and Huan Liu. Adaptive unsupervised feature selection on attributed networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 92–100, 2019.

[216] Jundong Li, Xia Hu, Jiliang Tang, and Huan Liu. Unsupervised streaming feature selection in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1041–1050, 2015.