Evaluation Methods and Measures for Causal Learning Algorithms

Lu Cheng , Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan , and Huan Liu , *Fellow, IEEE*

Abstract—The convenient access to copious multifaceted data has encouraged machine learning researchers to reconsider correlation-based learning and embrace the opportunity of causality-based learning, i.e., causal machine learning (causal learning). Recent years have, therefore, witnessed great effort in developing causal learning algorithms aiming to help artificial intelligence (AI) achieve human-level intelligence. Due to the lack of ground-truth data, one of the biggest challenges in current causal learning research is algorithm evaluations. This largely impedes the cross-pollination of AI and causal inference and hinders the two fields to benefit from the advances of the other. To bridge from conventional causal inference (i.e., based on statistical methods) to causal learning with Big Data (i.e., the intersection of causal inference and machine learning), in this survey, we review commonly used datasets, evaluation methods, and measures for causal learning using an evaluation pipeline similar to conventional machine learning. We focus on the two fundamental causal inference tasks and causality-aware machine learning tasks. Limitations of current evaluation procedures are also discussed. We, then, examine popular causal inference tools/packages and conclude with primary challenges and opportunities for benchmarking causal learning algorithms in the era of Big Data. The survey seeks to bring to the forefront the urgency of developing publicly available benchmarks and consensus-building standards for causal learning evaluation with observational data. In doing so, we hope to broaden the discussions and facilitate collaboration to advance the innovation and application of causal learning.

Impact Statement—Causal learning goes beyond machine learning due to its power of uncovering data generating processes. Causality relates to crucial open problems in machine learning. On the opposite, machine learning contributes to addressing fundamental challenges in causal inference. One key challenge of causal learning is that the research domain lacks public benchmark resources to support principled evaluation of research contributions. Our goal is to promote objectivity, reproducibility, fairness, collaboration, and awareness of bias in causal learning research. Arguing that this goal can only be achieved through systematic, objective, and transparent evaluation, in this article, we provide a comprehensive review of the evaluation of fundamental tasks in causal inference and causality-aware machine learning tasks. Similar to the evaluation in conventional machine learning, the causal evaluation pipeline includes the evaluation protocols, metrics, datasets, and popular causal tools/packages.

Index Terms—Benchmarking, Big Data, causal inference, causal learning, evaluation.

I. INTRODUCTION

MACHINE learning (ML) is a key pillar of artificial intelligence (AI). The unmatched availability of Big Data unleashed the unprecedented power of ML to support situational awareness and decision making. While witnessing the exceptional success of ML technologies in various applications, users started to notice a critical shortcoming of ML. Traditional ML techniques can learn correlation-based patterns and relationships from data. *Unfortunately, correlation is a poor substitute for causality as in many cases, data may contain spurious correlations* [1], [2]. Causal inference with observational data—data that have been generated by something other than randomized experiments—offers a promising alternative to correlation-based learning. Causal inference excels at uncovering the underlying causal mechanisms, leading to its wide applications in a myriad of high-impact domains, such as education [3]–[6], medical science [7], [8], economics [9], epidemiology [10], [11], meteorology [12], and environmental health [13]. Therefore, learning causality is significant for AI achieving human-level intelligence [14].

The conventional way to understand causality is to use interventions and/or randomized controlled trials (RCTs) [15], [16]. In many situations, however, these are time consuming, impractical, or sometimes even unethical [16]–[18]. Attention has been drawn to the recent availability of big observational data in all walks of life as they provide new opportunities for learning causality, without the disadvantages of RCTs. While being relatively recent, causal ML [causal learning (CL)] with observational data is emerging as a vibrant field with new opportunities and domain specific challenges. CL "seeks to model the effect of interventions and distribution changes with a combination of data-driven learning and assumptions not already included in the statistical description of system" [19]. Note that causal inference is relevant to methods that perform inference (effect estimation) and structural learning (causal discovery), whereas CL is more frequently mentioned along with methods that leverage ML to improve causal inference tasks or use

Ruocheng Guo is with the School of Data Science, City University of Hong Kong, Hong Kong (e-mail: ruocheng.guo@cityu.edu.hk).

Lu Cheng, Raha Moraffah, Paras Sheth, K. Selçuk Candan, and Huan Liu are with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: lcheng35@asu.edu; rmoraffa@asu.edu; psheth5@asu.edu; candan@asu.edu; huanliu@asu.edu).

TABLE I
SUMMARY OF METRICS, PROCEDURES, DATASETS FOR EVALUATING CL APPROACHES

		Causal Effect Estimation	Causal Structure Learning	Causal Interpretability and Fairness		Unbiased Interactive ML
Metrics	**Standard Effect Metrics**	MAE, MSE, RMSE, PEHE, Policy Risk	SHD, SID, Frobenius Norm, Precision, Recall, F1, TPR, FPR, MSE, AUC, Precision-Recall Curve, FPR-TPR Curve, TVD, KL-Divergence, F-test	**Counterfactual Explanation**	Sparsity, Interpretability, Speed, Proximity, Diversity, Visual Linguistic	NDCG@K, MAP@K, ARP@K, APLT@K
	Heterogenous Effect Metrics	$Uplift_{Coef}$, $Qini_{Coef}$		**Fairness**	FACE, FACT, Counterfactual Fairness, PC-Fairness, Ctf-DE, Ctf-IE, Ctf-SE	
	Time Series Metrics	Standard and Heterogeneous Effect Metrics, F-Test, T-Test				
Procedures	**With Ground Truth**	Observational data with known effect; observational and experimental data pairs; sampling from observational data; sampling from synthetic data; sampling from RCTs	A transductive setting where we have the ground-truth causal graph and estimated graph	**Transductive**	Training on a regular dataset and testing on generated counterfactuals	Training set comes from a biased source whereas test set comes from an unbiased source
	Without Ground Truth	Evaluation is possible if subset of the data is from RCTs		**Inductive**	Generating counterfactual explanations for an unseen instance	
Dataset		Under Unconfoundedness Assumption, Natural Experiments, RCTs	Causal Direction, Causal Graphs, Time Series Datasets	Image, Text, Tabular		Semi-Synthetic datasets, RCTs

causality to resolve some limitations of current ML methods. In this article, we focus on the two fundamental problems in causal inference: 1) *learning causal effects (i.e., estimating the causal effect of a treatment on an important outcome)* and 2) *learning causal structure (i.e., examining whether a certain set of causal relations exists between variables)* [20]. Answering these challenging causal questions is starting to become feasible as large datasets may contain sufficient samples from the joint distribution of the observed variables [15], [21].

A primary obstacle that impedes research developments in CL is the lack of public benchmark resources to support principled evaluation. Standardized evaluation played a major role in the early days of ML research. Successful early benchmarking efforts, such as UCI ML[1] and UCI KDD[2] repositories, not only helped guide the development of efficient and effective ML algorithms, but also encouraged collaborative research and paved the way for the recent breakthroughs in deep learning. Unfortunately, given the different learning objectives of conventional ML and CL, these existing benchmarks are not applicable to CL.

The overarching goal is, therefore, to enable the advancement of CL research and promote objectivity, reproducibility, fairness, collaboration, and awareness of bias in CL research. We argue that this goal can only be achieved through systematic, objective, and transparent evaluation of CL models and algorithms. In this survey, we aim to provide a comprehensive review for the evaluations of the two fundamental tasks in causal inference and causality-aware ML tasks, such as causal interpretability. We summarize the results in Table I. Similar to evaluation in conventional ML, the evaluation pipeline for CL includes the evaluation protocols, metrics, datasets, and popular benchmarking tools/packages. Under each task, we, then, discuss the

limitations of current evaluation procedures. This survey also aims to help set an agenda in future research on benchmarking CL algorithms by examining prominent open problems and challenges in current evaluation pipeline for CL.

Difference from previous work: Here, we first briefly summarize the existing surveys on causal inference and CL and then highlight their differences from this article.

Guo *et al.* [16] focus on reviewing the methodologies for causal effect estimation and causal structure learning, as well as discussing the connections between causal inference and ML. Yao *et al.* [22] specifically survey causal effect estimation methods and tools under the potential outcome framework. Another survey by Spirtes and Zhang [23] reviews semiparametric-score-based methods for the learning causal structure with independent and identically distributed (i.i.d.) and time-series observational data. To bridge from ML to artificial general intelligence, Schölkopf [20] shows that many challenging problems in ML and AI are inherently related to causality. The author especially examines where the links between AI and graphical causal inference have been and should be established. The rapid growth of time-series data and the unique challenges it brings in causal studies have led to surveys such as [24] reviewing problems, methods, and evaluation related to causal time-series analysis. In causal ML, Chen *et al.* [25] summarize different types of bias in recommendation systems and review methods that aim to mitigate such bias by leveraging causal inference theories. Moraffah *et al.* [26] survey the methods that leverage causality to enable the interpretability of ML models. All previously mentioned surveys center on reviewing the theories and methods in CL without an in-depth discussion about the evaluation.

The most related work is by Shimoni *et al.* [27]. The former introduces a benchmarking framework for causal effect estimation. It summarizes evaluation metrics and data generation with ground-truth effect. The second work surveys methods and benchmarks for causal structure learning. Compared to prior

studies, this survey aims to present a comprehensive review of the evaluation protocols, *evaluation metrics*, datasets, and causal tools/packages that have been widely used to benchmark causal effect estimation and causal structure learning. In addition, with the growing interest in causality in AI community, we also examine the existing framework for evaluating causality-aware ML tasks via several representative examples such as causal interpretability and fairness. Therefore, this survey complements the existing surveys focused on 1) reviewing causal theories and methodologies and 2) benchmarking one fundamental task in causal inference, i.e., either causal structure learning or causal effect estimation.

Intended audience and article organization: This survey will most benefit researchers and practitioners who have basic knowledge of causal inference and would like to develop or apply CL algorithms, but often find the evaluation rather challenging. It will also be useful for those who are interested in knowing the differences between evaluating standard ML and CL algorithms.

The rest of this article is organized as follows. We discuss evaluation frameworks for causal effect estimation in Section II, causal structure learning in Section III, and causal ML tasks in Section IV. We, then, summarize and compare the existing tools and packages for causal inference in Section V. Section VI delineates prominent open problems and challenges. Finally, Section VII concludes this article.

II. BENCHMARKING CAUSAL EFFECT ESTIMATION

We specify the pipelines for evaluating the first fundamental task in causal inference—causal effect estimation. We begin by defining the problem in the Neyman/Rubin Potential Outcome Framework [28] and, then, introduce the benchmarking datasets, evaluation procedures, and metrics.

Many empirical analyses are motivated by the necessity to estimate the causal effects of a binary treatment on the outcome of interest. Let $\boldsymbol{x}_i$, $t_i \in \{0, 1\}$, and y_i be the features (i.e., background variables), treatment assignment with $t_i = 1$ being under treated and $t_i = 0$ under control, and outcome of unit i. Causal effect estimation is defined as follows.

Definition 1 (Causal effect estimation): Given n units $\{(\boldsymbol{x}_1, t_1, y_1), \ldots, (\boldsymbol{x}_n, t_n, y_n)\}$, estimating causal effects is to quantify the changes of Y as we alter the treatment assignment from 0 to 1.

Under different contexts, effects can be estimated within the entire population, a subpopulation that is defined by background variables, an unknown subpopulation, or an individual. An average treatment effect (ATE) τ is typically used for assessing a population represented by the distribution of X

$$\tau = \mathbb{E}_X[\tau(\boldsymbol{x})] = \mathbb{E}_X[Y|do(T = 1)] - \mathbb{E}_X[Y|do(T = 0)] \tag{1}$$

where $do(T = 1)$ and $do(T = 0)$ indicate that the treatment assignment is "treated" and "control," respectively. When the population is heterogeneous, the ATE can be misleading as the same treatment may affect individuals differently. Under heterogeneity, a common assumption requires that each subpopulation

is defined by a set of features, i.e., conditional ATE (CATE)

$$\text{CATE}: \tau(\boldsymbol{x}) = \mathbb{E}[Y|do(t = 1), \boldsymbol{x}] - \mathbb{E}[Y|do(t = 0), \boldsymbol{x}]. \tag{2}$$

An individual treatment effect (ITE) is a contrast between potential outcomes of a unit

$$\tau_i = Y_i(do(t = 1)) - Y_i(do(t = 0)). \tag{3}$$

Note that τ_i is not necessarily equal to $\tau(\boldsymbol{x})$ as the latter is an average over a subpopulation. The goal of the *causal effect estimation* task is to learn a function $\hat{\tau}$ that estimates the ATE or the CATE, depending on the degree of homogeneity of the population, for binary treatment options: $T = 0$ and $T = 1$.

A. Evaluation Metrics

Evaluation metrics for causal effect estimation can be categorized into metrics for standard and heterogeneous effect estimations, as well as for time-series effect estimation. We review popular metrics in each category.

1) Metrics for Standard Causal Effect Estimation: As effect is typically continuous, most of the metrics are directly adapted from those for regression in ML, e.g., root-mean-square error (RMSE), mean absolute error (MAE), precision in estimation of heterogeneous effect (PEHE), and policy risk. Given the ground-truth ATE τ and the predicted ATE $\hat{\tau}$, the MAE of estimating the ATE is defined as

$$\epsilon_{\text{MAE\_ATE}} = \frac{1}{M} \sum_{j=1}^{M} |\tau_j - \hat{\tau}_j| \tag{4}$$

where M stands for the number of experiments and j is the index of the experiment. There are other similar metrics, such as *mean squared error* (MSE) and RMSE [29], [30]

$$\epsilon_{\text{MSE\_ATE}} = \frac{1}{M} \sum_{j=1}^{M} (\tau_j - \hat{\tau}_j)^2$$

$$\epsilon_{\text{RMSE\_ATE}} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} (\tau_j - \hat{\tau}_j)^2}. \tag{5}$$

The PEHE is used for estimating the CATE and is defined as

$$\epsilon_{\text{PEHE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i^1 - y_i^0 - \hat{\tau}(\boldsymbol{x}_i))^2 \tag{6}$$

where y_i^1 and y_i^0 are the two potential outcomes of the ith unit, $y_i^1 - y_i^0$ denotes the ground-truth CATE, and $\hat{\tau}(\cdot)$ is the learned function. When the true ITE is unknown but a subset of the dataset comes from an RCT, one can use

Policy risk, defined as the average loss in value when treatment is assigned based on the ITE estimator-guided policy [31], [32]

$$\text{Policy risk} = 1 - \left(\mathbb{E}[\tilde{y}_i^1 | t_i = 1, \pi_i = 1] p(\pi_i = 1) \right.$$

$$\left. + \mathbb{E}[\tilde{y}_i^0 | t_i = 0, \pi_i = 0] p(\pi_i = 0) \right) \tag{7}$$

where $\pi_i = 1$ denotes the policy to treat, and to not treat $\pi_i = 0$ otherwise. $\tilde{y}_i$ is the factual outcome scaled between [0,1]. The second component in (7) represents the expected potential

outcome, i.e., the weighted sum of expectations of two potential outcomes.

2) Metrics for Heterogeneous Effect Estimation: Evaluating heterogeneous effects with binary treatment typically follows the uplift modeling literature [33], [34]. Uplift modeling is an ML approach that employs the potential outcome framework to estimate the ITE in order to customize treatment assignments for different units. For example, companies use uplift models to differ between customers who buy product because of the campaign and those who will buy anyway. The evaluation metrics are defined over a curve measuring the performance of algorithms: the x-axis of the curve represents the number of units or the percentile of the population sorted by each estimated ITE (the larger the estimated ITE, the smaller the percentile); the y-axis represents gain when the treatment is assigned to the top a-percentile. Given population sorted by the inferred ITEs, *uplift curve* (also known as cumulative gain chart) measures the average cumulative gain of receiving the treatment in the first b units

$$\text{Uplift}(b) = \left(\frac{Y_b^1}{N_b^1} - \frac{Y_b^0}{N_b^0} \right) (N_b^1 + N_b^0) \tag{8}$$

where Y_b^1 (Y_b^0) and N_b^1 (N_b^0) denote the sum of the treated (control) outcomes and the number of treated (control) units among the first b units, respectively. A variant of uplift curve is the *Qini curve* that is defined in terms of percentiles, instead of the absolute number of units [35]

$$\text{Qini}(a) = Y_a^1 - \frac{Y_a^0 N_a^1}{N_a^0} \tag{9}$$

where Y_a^1 (Y_a^0) and N_a^1 (N_a^0) denote the sum of the treated (control) outcomes and the number of treated (control) units in the first a percentile of population, respectively. Similar to the area under the curve (AUC)–receiver operating characteristics (ROC) curve, we can compute the area under the Uplift/Qini curve, referred to as the uplift/Qini coefficient

$$\text{Uplift}_{\text{Coef}} = \sum_{b=0}^{N-1} \frac{1}{2}(\text{Uplift}(b+1) + \text{Uplift}(b))$$

$$\text{Qini}_{\text{Coef}} = \sum_{a=0}^{N-1} \frac{1}{2}(\text{Qini}(a+1) + \text{Qini}(a)) \tag{10}$$

where N is the size of the population.

3) Metrics for Time-Series Effect Estimation: Time-series data are sequences of real-valued data ordered over time. Causal inference for time-series analysis is of particular interest because many scientific questions involve causal effect estimation and causal structure learning using chronological observations [24], such as those related to medicine and social science. For example, social scientists have been interested in studying the effects of minimum wages on employment, where monthly employment rate needs to be recorded before and after a change in the minimum wage is made [36].

Evaluation metrics for standard effect estimation (such as MAE described above) can be used for time-series effect estimation at each time step s. To evaluate over the entire time

series, we need to further take an average of the results over time, e.g., $\text{MAE} = \sum_{s=1}^{S} \text{MAE}_s$. In addition, we can use metrics specifically designed for sequential data such as F-test and T-test. F-test assesses treatment effect heterogeneity by comparing the marginal variances of two potential outcomes. Let $\hat{e}_t$ and $\hat{u}_t$ be the time-dependent errors for the treated and control groups, respectively. p denotes the time lag; F-test is defined as

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/p}{\text{RSS}_1/(S - 2p - 1)} \tag{11}$$

where $\text{RSS}_0 = \sum_{s=1}^{S} \hat{e}_t$ and $\text{RSS}_1 = \sum_{s=1}^{S} \hat{u}_t$. To test the significance of a cause, one can also use the unpaired t-test (UTT) to compare the sequence of the treated group with that of the control group

$$\text{UTT} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1 - 1 + n_2 - 1)} \right)}} \tag{12}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the mean of two sequences, s_1 and s_2 denote the standard deviations, and n_1 and n_2 denote the cardinalities of the sequences.

B. Evaluation Procedures

Evaluating causal inference methods is significantly more challenging than evaluating purely associational methods. The primary obstacle in the evaluation procedure using observational data is that we typically cannot know the true causal effects. Some prior work uses observational data with known treatment effect. However, this requires that the studied phenomena are so well understood that the causal effect is obvious [37], limiting data availability. Another strategy uses data from pairs of observational and experimental studies to create a nearly ideal scenario for causal effect estimation with observational data [38]. This still suffers from low data availability. The most straightforward approach is to generate observational data from synthetic causal systems, where the treatment effect is either directly known or can be easily derived from the formulation [39], [40]. In particular, consider a data generating process that produces binary treatment T, potential outcome $Y(t)$, and multiple covariates $X = \{X_1, X_2, \ldots, X_k\}$. For each unit i, both potential outcomes $Y_i(1)$ and $Y_i(0)$ are measured. Given the biasing covariates $X^b \subseteq X$ of a unit and the synthetic dataset, one can begin by generating the selected treatment T^s using the biasing covariates $P(T^s|X^b) := f(X^b)$. If T^s is the same as the unit's corresponding treatment in the synthetic data, this unit is, then, added to the simulated observational dataset. This evaluation procedure is *sampling from synthetic data*. However, synthetic data cannot generalize well to real-world settings.

A more recent class of existing work augments an observational study with synthetic treatment assignment and outcomes generated by a synthetic function [27]. Given the observed covariates, we first randomly select a subset of samples, which is, then, fed into a data generating process to simulate treatment assignment and potential outcomes. Causal models are, then, trained on the tuples of treatment, covariates, and factual outcome. This approach—referred to as *sampling from*

observational data—can be used to evaluate both individual- and population-level effects. The newest approach creates constructed observational data by *sampling from RCTs* [41]. It has been shown that, in expectation, this approach can create datasets equivalent to those produced by randomly sampling from empirical datasets, where all potential outcomes are known. This approach is only suited for estimating population-level effects. Once the ground truth is available, evaluating causal effect estimation is similar to supervised learning algorithms. Without ground truth, evaluation is still possible if the subset of data is from RCTs, e.g., policy risk for the ITE.

The evaluation of causal effect estimation with time-series and networked data typically follows the same protocol and metrics as that with i.i.d. data. The estimated ATE (ITE) of the test set will be compared with the corresponding ground truth using MAE and MSE (PEHE).

C. Benchmarking Datasets

In general, there are two types of data used for identifying and estimating causal effects [16]: 1) data used under the *unconfoundedness* assumption [15]—observed/measured variables are sufficient to capture the causal links between treatment and outcome and 2) data collected from natural experiments, popular alternatives to RCTs. In the following, we briefly introduce exemplar benchmarking datasets in each category. For a comprehensive description, refer to [18]. We first introduce data used under the unconfoundedness assumption.

1) Datasets With Binary Treatment Variables: Most of the methods in the literature of learning causal effects are evaluated on datasets with binary treatment.

a) *Jobs:* The research studies the effect of a job training program on the real earnings of an individual. This dataset consists of RCT conducted by LaLonde [3] (297 treated and 425 control) and the Panel Study of Income Dynamics comparison group (2490 control) [42]. Features are multiple demographic variables such as age and race.

b) *Infant Health and Development Program (IHDP):* This is a dataset with simulated treatment and outcomes, initially complied by [6], seeking to evaluate the efficacy of comprehensive early intervention in reducing the development and health problems of low birth weight and premature infants. A commonly used simulation setting is the setting "A" in the NPCI package.[3] This dataset comprises 747 instances (139 treated and 608 control); each is associated with 25 features.

c) *Atlantic Causal Inference Conference Benchmark [43]:* It inherits the same features as those in IHDP data [44]. Various settings have been adopted to synthesize the treatment and outcomes.

d) *Twins:* The Twins' dataset in [45] is used to study the ITE of twins' weights on their mortality in the first year of their births. Each twin pair is represented by 46 features relating to the parents, the pregnancy, and the birth.
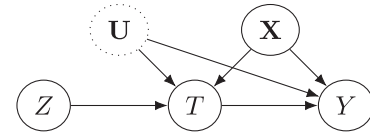

Fig. 1.   Causal graph of a valid IV (Z) when there are hidden confounders ($\mathbf{U}$). $\mathbf{X}$ stands for the observed confounders, and $\mathbf{U}$ is a set of hidden confounders.

e) *BlogCatalog* is used to study causal effect estimation with networked data. It is collected from an online social network service, where users can post blogs. Treatments and outcomes are synthesized based on the observed features, the social network structures, and the homophily phenomenon. This dataset has 5196 instances, 173 468 edges, and 8189 observed features.

f) *Amazon* [32] is an extension of the dataset in [46]. The goal is to evaluate the efficacy of positive (or negative) reviews in promoting product sales. Treatment and potential outcomes are generated based on the ratings and by matching the most similar products with different treatment assignments.

2) Datasets With Multiple Treatments/Continuous Treatment: There are studies in which multiple treatments are studied, e.g., *Twins-Mult* [47], *TCGA* [48], and *News-Mult* [48]. A dataset with continuous treatment can be found in [49]. It was collected to study the causal effect of the amount of smoking on the medical expenditure.

Datasets collected from natural experiments[4] allow us to relax the stringent unconfoundedness assumption [50] as it considers the presence of hidden confounders. Such datasets include variables selected under causal knowledge, such as the *instrumental variables (IVs)* and *running variable(s)*. In the following, we review the datasets with IVs [51] and regression discontinuity design (RDD) [52]. We also discuss datasets for causal time-series analysis and datasets with network information given the commonness of these two types of data.

3) Datasets with IVs: The IV is a powerful tool to identify the causal effect when there exist hidden confounders. We first briefly describe the core idea of the IV and, then, describe exemplar datasets.

Definition 2 (Instrumental variable): Given an observed variable Z, observed features $\mathbf{X}$, treatment T, and outcome Y, Z is a valid IV for the causal effect of $T \rightarrow Y$ iff Z satisfies: 1) $Z \not\!\perp\!\!\!\perp T | \mathbf{X}$ and 2) $Z \perp\!\!\!\perp Y | \mathbf{X}, do(T)$ [51].

The definition indicates that a valid IV causally influences the outcome only through the treatment. As shown in a causal graph of a valid IV (Z) in Fig. 1, the first condition requires that there is an edge $Z \rightarrow T$ or a nonempty set of colliders $\mathbf{X}$ s.t. $Z \rightarrow T \leftarrow \mathbf{X}$, where $\mathbf{X}$ denotes the observed confounders. The second condition requires that $Z \rightarrow T \rightarrow Y$ is the only path that starts from Z and ends at Y. Therefore, blocking T renders $Z \perp\!\!\!\perp Y$. This implies the *exclusive restriction* that there is no

---

[3][Online]. Available: https://github.com/vdorie/npci

[4]In most of the cases, natural experiments only allow us to identify ATEs but not ITEs or CATEs; therefore, we can use the ATE metrics, such as $\epsilon_{\text{MAE\_ATE}}$, $\epsilon_{\text{MSE\_ATE}}$, and $\epsilon_{\text{RMSE\_ATE}}$ for evaluation [29] and [30].

direct edge $Z \rightarrow Y$ or path $Z \rightarrow \mathbf{X}' \rightarrow Y$, where $\mathbf{X}' \subseteq \mathbf{X}$. Formally, $Y(do(Z = z), T) = Y(do(Z = z'), T) \; \forall T, z \neq z'$. We denote a dataset with the IV as $(\mathbf{x}_i, t_i, y_i, z_i)_{i=1}^{N}$, where z_i represents the IV of the ith instance. We can, then, evaluate approaches that leverage the IV to estimate the ATE, e.g., the ratio estimator [53]. Exemplar datasets are introduced as follows.

1) *1980 census extract:* This dataset contains 329 509 observations on the following variables: log weekly wage, quarter of birth (1–4), year of birth (30–39), place of birth (1980 census state codes), and education (highest grade completed). The goal is to study the causal effect of education on earning. The quarter of birth is considered as a legitimate IV [54] for years of schooling. This can be justified by considering that children born earlier in the year enter school at an older age and are allowed to drop out (on their 16th or 17th birthday) after having completed less schooling than those born later in the year.

2) *Current population survey extract:* This dataset contains 30 967 instances. Each instance represents a male born in 1944–1953 extracted from the 1979 and 1981–1985 March Current Population Survey. Each instance is matched with a dummy variable, which takes value from 25 lottery numbers. There are in total 72 variables. A potentially valid IV is the lottery number that represents the likelihood of serving in the military during the Vietnam era. This is because serving the military influences both schooling and earnings of a male and is determined by the lottery number. In addition, the lottery number is randomly assigned to each individual.

3) *Slave export and trust in community and society:* This study evaluates the long-term causal effect of slave export on trust in the community and society [55]. In [30], the treatment is defined as $t = \ln(1 + \frac{\text{slave-export}}{\text{area}})$ and the outcome is the trust in neighbors. The dataset has 6932 instances and 59 features. The IV used in [30] and [55] is each community's distance to the sea.

4) Datasets for RDD: RDD is a type of natural experiment, which has been found useful in a variety of real-world problems [52]. A dataset for RDD is denoted by $(\mathbf{x}_i, t_i, y_i, r_i)_{i=1}^{N}$, where r_i is the running variable of the ith instance that determines the treatment assignment of i based on a predefined cutoff value r_0. Running variable is used to control for confounding bias. For example, in the study of the causal effect of drinking alcohol on youth mortality [56], age is a natural running variable and $r_0 = 21$ (years old). Here, we present a list of datasets for RDD methods.

a) *Population threshold dataset [57]:* Decision making on many features of municipal governments (e.g., the electoral system, mayors' salaries, and the number of councillors) typically depends on whether the municipality is above or below arbitrary population thresholds. This dataset enables us to evaluate RDD methods for estimating ATEs of threshold-based policies on population-level political and economic outcomes.

b) *Pretest scores and posttest scores [58]:* It is used to study the causal effect of students' pretest scores on their posttest scores. This semisynthetic dataset is generated based on actual student test scores and demographic information. It has 2767 observations and uses pretest score as the running variable.

5) Datasets for Time-Series Effect Estimation: Here, we present a list of some commonly used real-world datasets for the time-series effect estimation. For a more comprehensive list of datasets, see [24].

a) *MIMIC II/III Data [59]:* These are large-scale data containing rich information relating to patients admitted to critical care units at a large tertiary care hospital. Some key covariates included are blood pressure, oxygen saturation, given medicine, and other temporal attributes such as time-stamped nurse-verified physiological measurements.

b) *Advertisement Data [60]:* These are used to measure the effect of an advertisement campaign on the number of times a user was directed to the advertiser's website from the Google search results page. Particularly, the product-related ads displayed alongside Google's search results for specific keywords went live for six consecutive weeks. The outcome variable includes search-related visits to the advertiser's website, such as organic clicks.

c) *Air Quality Data [61]:* These are used to study whether US gasoline content regulations successfully reduced ozone pollution. The data consist of two sources: 1) data on ambient air concentrations of ozone from the environmental protection agency's air quality standards database for 1989–2003 and 2) weather data measurements from the National Climatic Data Center's Cooperative Station Data (NOAA 2008), such as minimum and maximum temperatures.

6) Networked Data for Effect Estimation: Here, we present a list of the existing semisynthetic datasets for evaluating causal effect estimation with networked data [21], [62].

a) *BlogCatalog* [21] is a semisynthetic observational network dataset. It is based on real-world network structure and node attributes collected from the blog website BlogCatalog. Each node is a blogger and the node attributes are bag-of-words representation of her/his keywords. Each edge represents a friendship relation. The treatments and outcomes are sampled from a set of predefined structural equations, where treatment (control) means the blogger's content is viewed more by mobile (desktop) devices. Outcome is a real number standing for users' opinion on the bloggers' content. *Flickr* [21] is created in a similar way. Note that these two datasets did not consider interference.

b) *Wave 1* [62] is an in-school questionnaire dataset collected by Chantala and Tabor [63]. Ma and Tresp [62] creates a k-nearest neighbor (KNN) graph based on similarity between instances in the original Wave 1 data. The KNN network has 5578 nodes and 100 158 edges. Each node is a student and node attributes include age, grade, health insurance, and so on. Outcome is students' performance. Treatment is whether a student is assigned to a tutor program. Note that this dataset includes simulated peer effect to consider interference.

c) Pokec [62] is semisynthetic and is generated from a real-world social network dataset [64]. Each node is a user, and the node attributes include age, gender, education, and so on. Treatment means under exposure of a certain medicine advertisement, and outcome represents whether the user made a purchase of it. The purchase can be caused by either the treatment or the interference. Similar to Wave 1, the peer effect is simulated from a predefined structural equation.

D. Discussion

Revisiting current evaluation pipelines for causal effect estimation manifests that there still remains much to be done toward an effective benchmarking framework. The major challenge we confront when collecting data for causal effect estimation is how to design ethical, fast, reliable, and easy-to-implement experiments. RCTs are mostly impractical given the financial and ethical considerations. We also have limited access to the background variables useful for controlling for confounding bias. For example, users' social network information—a common hidden confounder in observational studies—can be used to approximate their socioeconomic status [21], [65]. However, collecting personal data from social networking sites may contradict the terms of user privacy [66]. In terms of effect estimation with networked data (e.g., social network information), in addition to the limited data challenge, extra effort is needed to examine the model robustness against the interference-related assumptions. Given that many outcomes of interest in the social and health sciences might be ordinal and do not have a meaningful scale, model evaluation requires careful treatment in these scenarios [67]. Moreover, collecting data with ground-truth ITEs is almost impossible in reality. This is because counterfactual outcomes cannot be observed at the individual level; otherwise, ITEs would not be needed. In time-series causal effect estimation, data collection is even more challenging because outcomes and time-varying covariates may take months or even years to observe [24], [68]. Popular alternatives use synthetic or semisynthetic datasets where treatments and outcomes are synthesized based on real-world data. Therefore, it is a crucial first step to develop high-quality simulation that mimic the real-world data generating processes for benchmarking causal inference algorithms. From the data perspective, comparatively less effort has been focused on data with auxiliary information under unconfoundedness assumption. Real-world data have rich auxiliary information (e.g., social media data) from multiple modalities and can overcome the limitations in using synthetic and semisynthetic datasets. Developing and evaluating causal models using such data is important. Another critical missing element in current evaluation procedures is how to validate models with observational data. There have been debates about whether cross-validation—the standard recipe for validating ML models—can be used for causal inference. Recently, it has been more commonly agreed that it is generally impossible to extend cross-validation in causal inference tasks because cross-validation is used to improve predictive accuracy and assure models' repeatability, but says nothing about causality.

Consequently, evaluation metrics and procedures for hyperparameter tuning and causal model selection are in great need. The overarching goal is to develop an open-source platform/software that provides objective, transparent, independent, and easy-to-use evaluation of causal algorithms.

III. Benchmarking Causal Structure Learning

Causal structure learning refers to the task of identifying the causal relations for a given set of variables $V = \{X_1, \ldots, X_n\}$. The goal is to generate a causal graph $G = \{V, E\}$ that represents the causal relations over the set of variables in V. E represents the set of directed edges between the variables in V. Causal relation between two variables X_i and X_j is defined as follows.

Definition 3 (Causal relation): For any directed edge e in G, $X_i \rightarrow X_j$ implies that X_i is a direct cause of X_j relative to variables in V.

Causal structure learning can then be defined as

Definition 4 (Learning causal structure): Given observed samples of J variables, $\{X_{,j}\}_{j=1}^{J}$, we aim to determine whether the value of the jth variable $X_{,j}$ would change significantly if we modify the value of the j'th variable $X_{,j'}$ for all $j \neq j'$.

For instance, learning the causal structure in previous graduate admission example might help answer the question: "*Does changing the sex of an applicant causally affect her/his application result?*" One can use causal graphs to identify the effect that would occur in the other variables when a variable X_i is introduced to an intervention, i.e., when the value of X_i is set to a fixed value. Once a variable is intervened on, a subgraph is generated by removing all the directed edges that point toward X_i in G. By repeating this process for all the variables in V, one can identify those causal pathways that are active in any experiment. Given observational data, algorithms for causal structure learning aim to discover a set of causal graphs as candidates [69] under the assumption that causality can be identified among statistical dependencies [70], [71]. Methods that recover causal dependence from time-series data have the benefit of temporal precedence and are often based on bivariate Granger causality tests [72], which are used to determine whether X causes Y or Y causes X (conditioned on a set of covariates). Graphical Granger methods use a series of Granger causality tests to determine the full structure among a set of variables.

A. Evaluation Metrics

Evaluating causal structure learning compares each of the learned causal graph candidates $\hat{G}$ with a ground truth G to determine if they belong to the same equivalence class.

Definition 5 (Equivalence class): Two causal graphs G and $\hat{G}$ belong to the same equivalence class iff each conditional independence of G is also implied by $\hat{G}$ and *vice versa*.

For example, Figs. 2 and 3 show two causal graphs that belong to the same equivalence class as they share the same set of conditional independence $\{X_{,2} \perp\!\!\!\perp X_{,3} | X_{,1}\}$.

Commonly used evaluation metrics for causal structure learning can be categorized into two major types: 1) graph-distance-based measures and 2) classification-based measures. For the

Fig. 2.   Causal graph with the conditional independence $X_{,2} \perp\!\!\!\perp X_{,3}|X_{,1}$.

Fig. 3.   Different causal graph with the same conditional independence $X_{,2} \perp\!\!\!\perp X_{,3}|X_{,1}$.

first category, we introduce structural hamming distance (SHD), structural intervention distance (SID), and Frobenius norm. A comprehensive study of metrics comparing learned causal graphs can be found in [73].

1) *SHD [74]–[76]:* Given two causal graphs, one being the ground-truth partially DAG[5] and the other being a predicted partially DAG, SHD is defined as the number of edits (adding, removing, or reversing an edge) that have to be made to the learned graph $\hat{G}$ for it to become the ground-truth graph G. It can be formulated as

$$\text{SHD} = A + D + I \qquad (13)$$

where A represents the number of edges that are added, D represents the number of edges that were deleted, and I represents wrongly oriented edges. Researchers generally present normalized SHD by calculating the ratio of different baselines' SHD over the SHD for their proposed methods.

2) *Frobenius norm* measures the difference between two adjacency matrices of two causal graphs [77]. Formally

$$\text{Frobenius norm}: \sqrt{\text{trace}\left\{\left(\mathbf{B}_{\text{true}} - \widehat{\mathbf{B}}\right)^T \left(\mathbf{B}_{\text{true}} - \widehat{\mathbf{B}}\right)\right\}} \qquad (14)$$

where $\mathbf{B}_{\text{true}}$ and $\widehat{\mathbf{B}}$ are the true and predicted adjacency matrices, respectively.

3) *SID:* For causal-structure-learning-based methods, it is important to understand the causal interpretations of a graph since it helps predict the result of interventions. Given a true DAG G and an estimated DAG $\hat{G}$, SID aims to infer the number of falsely inferred intervention distributions. SID is formulated as [74]

$$\text{SID} : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{N}$$

$(\mathcal{G}, \hat{\mathcal{G}}) \mapsto \#\{(i, j), i \neq j \mid \text{the intervention distribution}$

from i to j is falsely estimated by $\hat{G}$ with respect to $G\}$ $\qquad (15)$

where $\mathbb{G}$ is the space of DAGs defined over variables in V. SID is well suited to evaluate graphs for interventions.

The second type of metrics are based on the intuition that directional adjacency relations can be treated as a binary classification problem. Therefore, a variety of metrics in classification tasks can be used.

1) *Precision* is the ratio of true positives (TP) over sum of TP and false positives (FP), i.e., precision $= \frac{\text{TP}}{\text{TP}+\text{FP}}$.

2) *Recall* is defined as the ratio of TP over sum of TP and false negatives (FN), i.e., recall $= \frac{\text{TP}}{\text{TP}+\text{FN}}$.

3) *F1 score* is the harmonic mean of precision and recall of the learned structure as compared to true causal structure.

4) *False positive rate (FPR):* In terms of graph, FPR is defined as the ratio of the edges that are present in the predicted graph E_M but not present in the ground-truth graph E_{GT} over the absolute difference between ground-truth edges and all possible edges. It is formulated as

$$\text{FPR} = \sum_i \frac{e_i}{|E - E_{\text{GT}}|}, \quad e_i \in E_M \backslash E_{\text{GT}}. \qquad (16)$$

5) *True positive rate (TPR):* In terms of graph, TPR is defined as the ratio of the common edges between the ground-truth and predicted causal graphs over the number of edges in ground-truth graph. Formally

$$\text{TPR} = \sum_i \frac{e_i}{|E_{\text{GT}}|}, \quad e_i \in E_M \cap E_{\text{GT}}. \qquad (17)$$

6) *MSE* is defined as the sum of square of difference between the predicted and the ground-truth causal graphs divided by the total number of nodes. It is formulated as

$$\text{MSE} = \frac{1}{|T|} \sum (T - A)^2 \qquad (18)$$

where T is the predicted adjacency matrix and A is the ground-truth adjacency matrix.

7) *Area under ROC curve* is the area under the curve of recall versus FPR at different thresholds.

8) *Areas under the precision–recall and FPR–TPR curves* can also be used [75], [78] given that accuracy is measured in terms of precision and recall.

Finally, recent works, including [79] and [80], highlight that evaluating against structural measures only may limit the researchers learning from experimental studies of the model's performance, which, in turn, creates a gap between theory and practice. To overcome this limitation, they suggest several interventional measures that compare the learned distribution to the ground truth obtained through interventions. Now, we list common interventional measures.

1) *Total variation distance (TVD)* measures the distance between two probability distributions. Under TVD, the quality of an estimated distribution relative to a known distribution can be computed as

$$\text{TVD}_{P, \hat{P}, T=t}(O) = \frac{1}{2} \sum_{o \in \Omega(O)} |P(O = o \mid do(T = t))$$
$$- \hat{P}(O = o \mid do(T = t))| \qquad (19)$$

where P denotes the true interventional distribution, $\hat{P}$ denotes the estimated interventional distribution, and $\Omega(O)$ denotes the domain of O.

---

[5]$\mathcal{G}$ is called a partially DAG if there is no directed cycle, i.e., no pair (j, k) such that there are directed paths from j to k and from k to j.

2) *Kullback–Leibler (KL) divergence* measures how one probability distribution differs from others. Given the estimated interventional distribution $\hat{P}$ and true interventional distribution P, the KL divergence for continuous distributions can then be defined as

$$D_{\text{KL}}(P(x)\|\hat{P}(x)) = \int_{-\infty}^{\infty} P(x) \ln \frac{P(x)}{\hat{P}(x)} dx. \quad (20)$$

As for time-series data, aside from the aforementioned metrics, F-test is a popular metric used to evaluate causal structure learning approaches. In this approach, with the defined null hypothesis, parameters are estimated for both the restricted and the unrestricted models. An F-statistic is then computed using the residual sum of squares (RSS) of the two series, which is given by

$$\frac{(\text{RSS}_R - \text{RSS}_{UR})/p}{\text{RSS}_{UR}/(T - 2p - 1)} \sim F_{p,T-2p-1} \quad (21)$$

where T is the length of the time series, p is the number of lags, RSS_R is the RSS for restricted model, and RSS_{UR} is the RSS for the unrestricted model. For a more detailed review, readers can refer to [79] and [80].

B. Evaluation Procedures

The evaluation procedure for algorithms that learn causal relations from observational data often adopts a transductive setting, where the evaluation is based on the structural difference between the learned graph and the ground-truth graph. This is similar to the standard procedure to evaluate a supervised learning algorithm. Aside from evaluating against structural methods only, recent works emphasize the need for evaluating causal structure learning models against interventional measures. Unlike structural measures, interventional measures penalize edge misorientation errors proportionally to their effect on the estimation of interventional effect.

C. Benchmark Datasets

Datasets in this task can be considered in two categories: 1) the first category includes datasets for learning the causal direction between two variables and 2) the second category includes datasets for learning the underlying causal graph of a set of observed variables. Ideal datasets are collected from real-world scenarios and annotated by the experts in the corresponding fields to evaluate algorithms for learning causal relations. However, since it is extremely challenging to obtain such real-world ground-truth datasets for causal graphs, synthetic datasets are commonly used for benchmarking purposes. Popular datasets used for learning causal directions include *Tübingen cause–effect pairs*, *Alzheimer's Disease Neuroimaging Initiative (ADNI)* [81], *AntiCD3/CD28* [82], and *Note* [83]. Datasets for learning causal graphs are *lung cancer simple set (LUCAS)*, *lung cancer set with probes (LUCAP)*, and *random chordal graphs*. We briefly introduce these datasets.

1) Datasets for Learning Causal Directions: We detail exemplar datasets for learning causal relations between two variables as follows:

a) *Tübingen cause–effect pairs [37]:* It consists of real-world samples from cause–effect pairs and is collected across various subject areas.
b) *ADNI [81]:* Alzheimer's Disease Neuroimaging Initiative consists of 819 subjects who were enrolled at baseline and followed for 12 months using standard cognitive and functional measures typical of clinical trials.
c) *AntiCD3/CD28 [82]:* This is a dataset of causal protein-signal networks with 853 instances of multi-variable individual-cell data. Each variable stands for a biomolecule.
d) *Note [83]:* This is a time-series dataset for inferring the arrow (direction) of the time.
e) *Abalone [84]:* It contains 4177 samples and each sample has four attributes: sex, length, diameter, and height.

Datasets for Learning Causal Graphs:
a) *LUCAS and LUCAP:LUCAS* and *LUCAP* consist of data generated by predefined causal graphs with binary variables.
b) *Random chordal graphs:* This is a synthetic dataset generated using approach discussed in [85] and [86].
c) *CausalWorld:* The CausalWorld [87] dataset is a comprehensive robotic benchmark dataset for transfer learning. It provides a combinatorial family of tasks with common causal structures and underlying factors, allowing the user to intervene on the causal variables to determine the similarity level of different tasks.
d) *SynTReN:* SynTReN is a network generator that creates synthetic transcriptional regulatory networks and produces simulated gene expression data that approximates experimental data. Topologies are created by selecting subnetworks from previous regulatory networks. Results in [88] show that the simulated topologies are closer to the real biological networks when compared to different random graph models. Several user-definable parameters adjust the complexity of the resulting dataset w.r.t. the structure learning algorithms.
e) *Causality 4 Climate:* The Causality 4 Climate [89] dataset comes from a NeurIPS 2019 competition for causal structure learning on climate time series. The dataset consists of an extensive number of climate model-based time-series datasets with known causal ground-truth. These datasets incorporate the main challenges of causal structure learning in climate research.

Causal Structure Learning Datasets for Time Series:
a) *Temperature ozone data [37], [90], [91]:* This dataset consists of ozone and radiation levels, across 72 points in time, and 16 different places. The assumed ground truth is that the radiation has a causal effect on ozone.
b) *Neural activity dataset:* This dataset consists of real-time whole-brain imaging to record the neural activity of Caenorhabditis elegans. The dataset consists of 302 neurons and is generally used to identify which neurons are responsible for the movement.
c) *Traffic prediction dataset [92]:* This dataset contains four months' worth of sensor data from Los Angeles, California. 207 sensors are placed for collecting this data. The

location of each sensor in the form of GPS coordinates is also included in the dataset.

d) *US manufacturing growth data [93]:* This dataset consists of microeconomic data of growth rates of US manufacturing firms in terms of employment, sales, research & development expenditure, and operating income, for the years 1973–2004. It can be used to identify the causal variables that affect the growth rate of a firm.

D. Discussions

The transductive setting has been the norm in the literature of learning causal relations [69], [94] in part because the conventional setting of learning causal relations is similar to that of supervised learning. However, in observational studies, the ground-truth causal relations may not be used to train the algorithms even during the training phase. In addition, Guo *et al.* [16] pointed out that most of the existing algorithms can only discover the causal relations among variables that can be observed in the training data. Hence, it remains an open problem to develop causal inference algorithms that can be generalized to unseen variables, i.e., an inductive evaluation setting. Furthermore, the combination of observational and experimental data may provide with unique opportunities to identify a model that, under various assumptions, can extract some true external cause–effect relationships. There is a recent line of research [95], [96] focusing on learning causal relations from the combination of observational data and interventional data with some selected variables. This is promising as it may overcome the limitations of using pure observational or RCT data. To this end, learning causal relations with the combined data can be reduced to an easier problem where the goal is to identify a certain set of interventions.

IV. EVALUATION OF CAUSALITY-AWARE ML TASKS

As a critical ingredient for AI to achieve human-level intelligence, causal inference has found itself in the spotlight of Trustworthy ML and Socially Responsible AI research [66], [97]. In this section, we discuss the evaluation pipeline of two representative tasks: causal interpretability and fairness and unbiased interactive ML.

A. Causal Interpretability and Fairness

The surge of ML algorithms for decision making in critical fields, such as law-making and autonomous cars, has made it necessary for the decisions made by these models to be interpretable by humans [98]. Interpretability in ML is defined as "model's ability to explain or to present in understandable terms to a human" [99]. To ensure the reliability and transparency of such decisions, causal interpretability aims to explain them by answering the counterfactual questions such as "what decisions would have been made if they had been under alternative situations (e.g., being trained with different inputs [100], [101] or model components [102], [103])?" We refer to this as a model's *causal interpretability*. Therefore, causal interpretable models are human-friendly and aim to answer causal questions.

1) Problem Statement: There is no unified definition for the causal interpretability of an ML model [26]. In this survey, we use a common definition, in which a model's causal interpretability is judged by the *counterfactual explanations* it generates for a set of inputs. Note that we skip the model-level counterfactual explanations in this article due to the lack of proper evaluation metrics and benchmarks for such frameworks in the literature [26].

Definition 6 (Example-level counterfactual explanation): Given an ML model and a predefined label, a counterfactual explanation of an instance is defined as a new instance that is generated by performing minimal changes to an original instance's features, which makes the model to predict the predefined label.

For example, when a person's credit card application fails, she may be interested in explanations that specify what minimal changes can be made to her profile to pass the application. Note that while similar to the definition of adversarial example, the example-level counterfactual is designed to explain the decision of the model rather than attack it. Generally, example-level counterfactual explanations aim to answer questions such as "*Why does this model predict a specific label for an instance?*" or "*Was the ith feature of the instance which caused the model to predict this label?*" Such questions can be answered using counterfactual inference [14]: Counterfactual distributions are new type of conditional probabilities (e.g., $P(y_x|x,'y')$) that indicate how likely the outcome of an observed instance, i.e., y', would change given x'.

Fairness also relates to making models more transparent and interpretable [98]. ML frameworks used for decision making in critical domains such as law making are required to make fair decisions and not discriminate against specific groups of people [104]. Several frameworks have tried to combine causal inference with fairness and propose a criterion for making fair decisions [105], [106].

2) Evaluation Metrics: We here discuss common evaluation metrics for two primary tasks in causal interpretability—example-level counterfactual explanation and fairness.

a) Example-level counterfactual explanations: One of the big challenges of evaluating counterfactual explanations is the lack of ground truth in traditional ML datasets. A commonly used alternative is to gauge the quality of the counterfactual explanations generated by an ML model in terms of metrics that aim to quantify certain characteristics of these counterfactual explanations. Moraffah *et al.* [26] have classified these metrics based on the properties of the explanation they are designed to measure. Particularly, six metrics—sparsity, interpretability, proximity, speed, diversity, and visual-linguistic metrics—have been suggested to measure the quality of the generated counterfactual explanations from different perspectives.

Sparsity metric aims to measure the amount of perturbation added to the original input in order to transform it to its corresponding counterfactual explanation. The lower the amount of perturbation is, the better is the quality of the counterfactual. Interpretable counterfactual explanations are often close to the training data manifold [107]. To gauge the closeness of counterfactuals to the data manifold, *Interpretability* metric is utilized and are generally computed based on the reconstruction errors

TABLE II
SUMMARY OF EVALUATION METRICS FOR COUNTERFACTUAL EXPLANATIONS [26]

	Counterfactual Characteristic	Description
1	Sparsity	Measures the amount of perturbation used to generate the counterfactual example [107, 108]
2	Interpretability	Measures the closeness of the counterfactual example to data manifold [107]
3	Proximity	Measure the similarity of the counterfactual example to original sample [109]
4	Speed	Measures the pace of counterfactual generation [107]
5	Diversity	Measures the diversity of the generated counterfactuals [109]
6	Visual-Linguistic Counterfactuals	Measures the high positeveness/negativeness of the visual explanation. Measures the consistency of the linguistic explanation with it's visual counterpart [110]

of the counterfactuals [107]. *Proximity* is another counterfactual evaluation metric which evaluates how close the counterfactual explanations are to the original samples. Unlike *Interpretability* metric, Proximity is usually calculated by calculating the l_p distance between the original and counterfactual samples. Since most interpretability frameworks are designed for real-world applications, counterfactual generation approaches should be fast. *Speed* metric measures the speed of counterfactual generation approaches. Another characteristic of high-quality counterfactuals is their diversity. This means that counterfactuals generated for a given input should be different. To measure the quality of these counterfactual explanation frameworks, *Diversity* metric can be used. Finally, to measure the quality of multimodal counterfactual explanations such as visual-linguistic explanations, one needs to report the high positiveness/negativeness of the visual explanations as well as the consistency of the visual and linguistic explanations. Table II summarizes these metrics with corresponding descriptions. For more detailed explanations, refer to [26].

b) Fairness: Evaluation of causal fairness is challenging and typically done via assessing models' performance in reducing biases. In the following, we discuss the state-of-the-art (SOTA) causality-based notions for fairness. For more details, refer to [111].

Khademi *et al.* [112] propose two notations based on causal effect estimation, i.e., fair on average causal effect (FACE) and fair on average causal effect on the treated (FACT), under the potential outcome framework [28]. With FACT notion, a classifier is fair iff

$$\mathbb{E}[Y_i^{(a_1)} - Y_i^{(a_0)}] = 0 \qquad (22)$$

which is equivalent to calculating ATE. FACT metric, on the other hand, is based on average causal treatment effect on the treated group, i.e., ATT. Based on this definition, a classifier is fair iff

$$\mathbb{E}[Y_i^{(a_1)} - Y_i^{(a_0)} | A^i = a_0] = 0. \qquad (23)$$

Another mainstream fairness notion proposed by Kusner *et al.* [105] is coined counterfactual fairness, which states that a classifier is fair if under any background variable X, we have

$$P(y_{a_1}|X = x, A = a_0) = P(y_{a_0}|X = x, A = a_0) \qquad (24)$$

where A is a set of protected attributes, Y is the outcome, and $X = V \backslash \{A, Y\}$ denotes the remaining variables in the system. Due to the challenge of causal identification using observational data, Wu *et al.* [106] propose path-specific counterfactual fairness (PC-Fairness), a generalized causal fairness notion which covers various causality-based fairness notions by tuning its hyperparameters. A classifier achieves PC-Fairness iff

$$P(y_{a_1|\pi, a_0|\bar{\pi}}|O) - P(y_{a_0}|O) = 0 \qquad (25)$$

where Y is the outcome, π denotes causal paths, and O is a set of observed variables.

Built on causal graphs, Zhang and Bareinboim [113] suggest counterfactual direct (Ctf-DE), indirect (Ctf-IE), and spurious effect (Ctf-SE) fairness measures. These metrics essentially assess the transmission from cause to effect. They are defined as follows:

$$\text{Ctf-DE}_{x_0, x_1}(y|x) = P(y_{x_1, W_{x_0}}|x) - P(y_{x_0}|x) \qquad (26)$$

$$\text{Ctf-IE}_{x_0, x_1}(y|x) = P(y_{x_0, W_{x_1}}|x) - P(y_{x_0}|x) \qquad (27)$$

$$\text{Ctf-SE}_{x_0, x_1}(y|x) = P(y_{x_0}|x_1) - P(y|x_0) \qquad (28)$$

where $X = x_1$ is the intervention on $Y = y$, $X = x_0$ is the baseline value of the sensitive attribute, and W indicates the mediator. Despite recent success in utilizing causality to measure fairness, evaluation of such frameworks is still an open problem mainly due to the "impossibility of fairness" [114]: existing fairness notions appear to be internally consistent but often mutually incompatible with each other.

3) Evaluation Procedures: Counterfactual explanations belong to post-hoc interpretability and are generated after a black-box ML model is trained. The first step of the evaluation procedure is to train an ML model on a regular dataset such as Adult-income [115]. Then, we need to generate counterfactual explanations based on the trained model and the dataset. Finally, we evaluate the generated counterfactual explanations using the existing criteria such as those in Table II. The standard procedure in the existing work (e.g., [109]) is transductive. One extension

is its inductive counterpart, where we check whether a method can generate counterfactual explanations for an unseen instance from a hold-out set [116].

4) Benchmark Datasets: To evaluate the causal interpretability of black-box ML models, we need datasets with two components: 1) a dataset for training and testing an ML task and 2) a source of ground truth that allows us to evaluate the causal interpretability of the generated explanations. Ideally, this would be a set of ground-truth counterfactual explanations for a given data instance and an ML model. Such ground truth can be extremely expensive to acquire manually given the large number of data instances and ML models. Hence, the second component of ground truth is commonly replaced by a set of proximal metrics [109]. These metrics can be conveniently computed given the generated counterfactual explanations and the corresponding instances from the original dataset. In the following, we introduce common datasets used for interpretability. It is worth mentioning that most of these datasets are not specifically designed for evaluating causal interpretability. Therefore, they do not come with the ground truth of counterfactual explanations that capture the causal aspect of the model. However, the evaluation metrics can be leveraged as proximal measures to determine the quality of a generated counterfactual explanation. Counterfactual explanations are model agnostic and, therefore, can work on datasets for most ML tasks. We categorize these datasets as Image, Text, and Tabular data. Following are exemplar datasets in each category.

Image datasets:

1) *ImageNet (ILSVRC)* [117] is an image dataset organized according to the WordNet hierarchy. There are more than 100 000 *synonym sets* or *synset* in WordNet, majority of them are nouns (80 000+). The ImageNet team aims to provide on average 1000 images to illustrate each synset. Images of each concept are quality controlled and human annotated. The task is to predict which synsets an image belongs to.

2) *MNIST* [118] consists of images of handwritten digits. It has a training set of 60 000 examples and a test set of 10 000 examples. It is a subset of a larger set available from the National Institute of Standards and Technology. The digits have been size-normalized and centered in a fixed-size image.

3) *PASCAL VOC 2012* [119] contains 12 031 training and validation images and 1456 test images. Among the 12 031 training images, 2913 of them have associated ground-truth object regions split evenly into a segmentation *train-2012* and *val-2012* sets.

Text datasets:

1) *20 Newsgroup Dataset [120]:* This dataset is a collection of nearly 20 000 news documents and is partitioned (nearly) evenly across 20 different newsgroups. This dataset has become increasingly popular for evaluating text mining and natural language processing algorithms w.r.t. text classification and clustering tasks.

2) *IMDB [121]:* The IMDB dataset contains 25 000 training documents, 25 000 test documents, and 50 000 unlabeled documents. It is a dataset for binary sentiment classification task. The dataset consists of movie reviews retrieved from the website IMDB.[6] For labeled documents, they are sampled in a way such that there is a 1:1 ratio between negative and positive documents.

3) *Amazon reviews [122]:* This dataset contains product reviews and metadata from Amazon, including up to 142.8 million reviews spanning May 1996–July 2014. Particularly, it includes reviews (ratings, text, and helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). One of the tasks can be evaluated by this dataset is binary sentiment classification.

Tabular datasets: The UCI repository [123] provides a myriad tabular datasets used by the ML literature. Here, we list those that are widely used for evaluating causal interpretability.

1) *Adult-income [115]:* This dataset contains demographic, educational, and other information based on 1994 Census database and is available on the UCI ML repository [123]. This dataset is preprocessed to include eight features, namely, hours per week, education level, occupation, work class, race, age, marital status, and sex. The ML model's task is to classify whether an individual's income is over 50 000.

2) *German loan dataset [124]:* This dataset contains 1000 observations of loan applicants, including numeric, categorical, and ordinal attributes. The label indicates whether the application is successful.

3) *LendingClub[7]:* This dataset consists of five years of loan records (2007–2011) by LendingClub company. After the dataset is preprocessed, it contains eight features: employment years, annual income, number of open credit accounts, credit history, loan grade as decided by LendingClub, home ownership, purpose, and the state of residence in the U.S.

4) *COMPAS:* This is a dataset collected by ProPublica [125]. It contains information for analysis on recidivism decisions in the U.S. After preprocessing, this dataset contains the following five features: bail applicant's age, gender, race, prior count of offenses, and the degree of criminal charge.

B. Unbiased Interactive ML

Recent years have witnessed increasing importance of interactive ML models and systems such as search ranking and recommendation systems in both academia and industry. Training such interactive models and systems is different from standard prediction tasks because the supervision signals or labels—e.g., ratings, clicks, and purchases—come from users' online behavior, collected through interactions between users and an ML model. The fundamental challenge in learning interactive ML models arises from the fact that they are trained/optimized with

---

[6][Online]. Available: https://www.imdb.com/
[7][Online]. Available: https://www.lendingclub.com/info/download-data.action

biased log data, while the goal is to optimize the performance via online experiments where RCTs can be conducted [126], [127]. This mismatch between training and test data makes causality critical in understanding and mitigating various types of bias in the process of training interactive ML models. In this section, we illustrate the uses of CL in one of the extensively studied tasks in interactive ML—unbiased learning to rank. Another related task is debiasing recommendation systems [25].

1) Problem Statement: Unbiased learning to rank is formally defined as follows.

Definition 7 (Unbiased learning to rank): Given a search result page (SERP) of a query q in the search log, $\boldsymbol{X}^q \in \mathcal{R}^{m^q \times d}$, $\boldsymbol{y}^q \in \{1, \ldots, m^q\}^q$, and $\boldsymbol{c}^q \in \{0, 1\}^q$ denote the feature matrix, the vector of item indices, and the vector of implicit feedback, respectively. d is the number of features and m^q is the number of items shown in the SERP of query q. Let $\mathcal{Q}$ and $\mathcal{Q}'$ be the set of queries for an offline training set and that of a test set, respectively. $f : \mathcal{R}^d \to \mathcal{R}$ denotes a ranking system that maps the feature vector of an item to its ranking score. Given offline search log data $\{\boldsymbol{X}^q, \boldsymbol{y}^q, \boldsymbol{c}^q\}_{q \in \mathcal{Q}}$, the goal is to optimize a specific ranking metric (e.g., NDCG) in a test set $\{\boldsymbol{X}^q, \boldsymbol{r}^q\}_{q \in \mathcal{Q}'}$ where $\boldsymbol{r}^q$ denotes the vector of unbiased feedback. $\boldsymbol{r}^q$ can be collected from RCTs or manual labeling (e.g., relevance score).

In an RCT, we rank all the items randomly and collect feedback from users. We call such feedback *unbiased* as it is only causally influenced by users' preference. While with offline log data, a user's feedback on an item is influenced by the user's preference, the position of the item ranked by the logging policy—the existing ranking system that generated the log data, and other potential factors (e.g., context of the item [128]). In some cases, if we do not have access to unbiased implicit feedback from RCTs for the test set, we can also evaluate a ranking system based on relevance scores as in [129]–[131].

2) Evaluation Procedures: The evaluation procedure of unbiased learning to rank is mostly similar to that of a standard learning to rank task in information retrieval. The major difference is that the training set and the test set are from different sources: the former is a typically biased offline dataset, whereas the latter is a dataset with unbiased feedback (often collected through RCTs). Given a test set $\{\boldsymbol{X}^q, \boldsymbol{r}^q\}_{q \in \mathcal{Q}'}$ and the corresponding predicted ranking $\{\hat{\boldsymbol{y}}^q\}_{q \in \mathcal{Q}'}$ of a ranking system, we introduce the following two most popular metrics to evaluate its performance.

Normalized discounted cumulative gain (NDCG@K) [132] is defined as

$$\text{NDCG@K} = \frac{1}{\text{IDCG@K}} \sum_{i=1}^{K} \left(\frac{2^{r_i^q} - 1}{\log_2(i+1)} \right) \quad (29)$$

where K is a positive integer (e.g., 10 or 50) and varies by specific applications; $\text{IDCG@K} = \sum_{i=1}^{K} \frac{1}{\log_2(i+1)}$ is the normalizer to ensure that NDCG@K is in the range [0,1]. Note that an item's rank is determined by the prediction $\hat{\boldsymbol{y}}^q$ of a ranking system. NDCG@K is a weighted sum of a function of unbiased feedback (e.g., relevance score) for the items ranked in top-K positions. A larger NDCG@K indicates that the top-K ranked items are more relevant.

Mean average precision (MAP@K) is defined as

$$\text{MAP@K} = \frac{1}{K} \sum_{i=1}^{K} |\{j | r_j^q = 1, j = 1, \ldots, i\}| / i. \quad (30)$$

MAP@K describes how accurate a ranking system's ranked predictions are, on average, over a whole validation/test dataset. The primary difference between MAP@K and NDCG@K is that the former assumes binary relevance, i.e., either relevant or irrelevant, whereas the latter also takes continuous values. In addition, there are metrics proposed specifically for debiasing popularity bias of ranking algorithms. The first metric is *average recommended popularity (ARP)* [133], defined as

$$\text{ARP@K} = \frac{1}{|\mathcal{Q}'|} \sum_{q \in \mathcal{Q}'} \frac{\sum_{i=1}^{K} \phi_i^q}{K} \quad (31)$$

where ϕ_i^q denotes the popularity of the document or product ranked at position i in the SERP of query q. We can also focus on the documents or products that are least popular, namely the long-tail documents or products. This leads to the *average percentage of long tail items (APLT@K)* [133], defined as

$$\text{APLT@K} = \frac{1}{|\mathcal{Q}'|} \sum_{q \in \mathcal{Q}'} \frac{\sum_{i=1}^{K} 1(i \in \mathcal{LT})}{K} \quad (32)$$

where $1(\cdot)$ is an indicator function. $\mathcal{LT}$ denotes a set of long-tail documents or products predefined by their popularity.

3) Benchmark Datasets: Data collection in unbiased learning to rank confronts general challenges in CL tasks. The alternative seeks to create semisynthetic datasets using a search log dataset with ground-truth relevance scores for each query–item pairs [130], [131]. First, we train a ranking system (e.g., RankSVM [131]) on 1% of the original training dataset. We then generate a list of items ranked in order using the entire training dataset $\{\boldsymbol{X}^q, \boldsymbol{r}^q\}_{q \in \mathcal{Q}}$. The next step employs a click model (e.g., the position-based click model [134]) to simulate users' clicks. Specifically, the position-based model assumes $P(c_i^q = 1) = P(o_i^q = 1)P(z_i^q = 1)$, indicating that a click on the ith item can only be observed iff the item is examined ($o_i^q = 1$) and relevant ($z_i^q = 1$). The binary relevance label z_i^q is defined as a function of the relevance score r_i^q:

$$P(z_i^q = 1) = \epsilon + (1 - \epsilon) \frac{2^{r_i^q}}{2^4 - 1} \quad (33)$$

where ϵ denotes the click noise and is often set to 0.1 [130], [131]. The maximum value of r_i^q is 4. The probability of examination is defined as a function of the position-based examination probability $\boldsymbol{p}$ [130], [131]

$$P(o_i^q = 1) = (p_i)^\eta \quad (34)$$

where $\eta \geq 1$ is a hyperparameter set to 1 by default. One popular dataset used for evaluating unbiased learning to rank models is the *Yahoo! Learning to Rank Challenge Dataset*. It contains 29 921 queries and 710k items [135]. Each query–item pair is represented by a feature vector of 700 dimensions and associated with a ground-truth relevance score in $\{0, 1, 2, 3, 4\}$. Any search log dataset with ground-truth relevance scores or feedback collected through RCTs can be potentially used in this task.

V. CAUSAL INFERENCE TOOLS

In this section, we examine popular tools/packages for benchmarking causal inference, including tools for causal effect estimation—*CausalML, EconML, DoWhy CauseBox*, for causal structure learning—*CausalNex, CausalDiscovery, pcalg, bnlearn, TETRAD*, and for evaluation—*Causality-Benchmark, JustCause*.

1) *CausalML* [136] implements an array of uplift modeling and causal inference methods using ML algorithms. It provides a standard interface for users to estimate CATE or ITE based on experimental or observational data, without strong assumptions on the model form. CausalML currently supports Tree-based algorithms (e.g., Uplift random forests on KL divergence), Meta-learner algorithms (e.g., S-learner, T-learner), and IV algorithms (e.g., two-stage lease squares). Covered evaluation metrics include RMSE and MAE.

2) *EconML* [137] estimates heterogeneous treatment effects from observational data via ML. The goal of EconML is to combine SOTA ML techniques with econometrics to automatically solve complex causal inference problems. The supported estimation methods are in the intersection of econometrics and ML, including double ML, orthogonal random forests, meta-learners, doubly robust learners, orthogonal IV, and deep IV. It does not include evaluation metrics for effect estimation.

3) Informed by conventional ML libraries for prediction, *DoWhy* [138] provides a unified interface for causal inference methods under the two fundamental frameworks— graphical models and potential outcomes. First, DoWhy creates a causal graphical model for each problem to describe the causal assumptions. It then uses graph-based criteria and *do*-calculus to find all potential ways of identifying a desired causal effect. In the third step, it estimates causal effects based on the identified estimand. Finally, DoWhy validates an effect estimate from a causal estimator using multiple refutation methods such as adding unobserved common causes and bootstrap validation. DoWhy can also call external estimation methods such as EconML and CausalML.

4) *CausalNex*[8] uses Bayesian Networks to combine ML and domain expertise for causal structure learning and effects estimation. It deploys SOTA structure learning method DAGs with NO TEARS [139] to understand the causal relationships between variables. CausalNex allows users to learn optimal graph structure through: 1) encoding domain expertise and 2) learning from data via structure learning algorithms. Users can also conduct counterfactual analysis by introducing *do*-calculus.

5) *CausalDiscovery* [140] implements an end-to-end step-by-step pipeline approach for recovering the direct dependencies (the skeleton of the causal graph) and the casual relationships between variables. CausalDiscovery currently includes 17 algorithms for graph skeleton identification and 19 algorithms for causal directed graph prediction. The goal of CausalDiscovery is to learn both the causal graph and the associated causal mechanisms from the joint probability distribution of observational data. It also includes R-based algorithms to extend the toolkit with additional R packages. There are two types of supported algorithms in CausalDiscovery—graph and pairwise causal inference models. In addition to the standard evaluation metrics, CausalDiscovery also includes SHD and SID.

6) *Causality-Benchmark* [27] benchmark algorithms that estimate ATE and ITE of an intervention on the outcome of interest. It includes unlabeled data for prediction, labeled data for validation, and scoring algorithms for automatic evaluation of prediction algorithms based on different evaluation metrics. As ground-truth data of causal effect cannot be known for real-world treatment, Causality-Benchmark uses a simulation-based approach that leverages the existing covariates and creates a causal graph to determine treatment assignment and effect. It supports a variety of metrics for causal effect estimation.

7) *JustCause*[9] evaluates causal inference methods using common datasets, including Twins, IHDP, and IBM ACIC. It can also generate synthetic datasets with a generic but standardized approach used in [141]. The goal of JustCause is to provide a fair and just way to benchmark new methods against several baselines and the SOTA methods in causal effect estimation. The supported algorithms are doubly robust estimation, inverse propensity weighting, S-learner, and T-learner. Evaluation metrics implemented include PEHE score, MAE, EnoRMSE score, and Bias.

8) *CauseBox*[10] is a benchmarking suite developed in Python for treatment effect estimation methods that consists of an ensemble of five deep-learning-based SOTA methods and two tree based methods. All the methods are evaluated on benchmarking datasets, including IHDP and News datasets. The goal of this toolbox is to provide a unified platform to compare algorithms for different metrics including PEHE and Policy Risk.

9) *bnlearn*[11] [142] is a causal structure learning toolbox developed in the R language. The bnlearn package provides implementations of various structure learning algorithms including, but not limited to, PC, Grow-Shrink (GS), Incremental Association Markov Blanket (IAMB), Hybrid Parents and Children (HPC), and Hill Climbing (HC). bnlearn also provides the conditional independence tests and network scores used to construct the Bayesian network. Both discrete and continuous data are supported. Furthermore, bnlearn facilitates choosing the learning algorithms based on different statistical criteria, so that the best combination for the data could be utilized.

---

[8][Online]. Available: https://github.com/quantumblacklabs/causalnex

[9][Online]. Available: https://github.com/inovex/justcause

[10][Online]. Available: https://github.com/paras2612/CauseBox

[11][Online]. Available: https://www.bnlearn.com/

TABLE III
COMPARISONS OF CAUSAL INFERENCE TOOLS WITH A FOCUS ON THE INCLUDED DATASETS, METHODS, AND METRICS

	Task	Causal Effect Estimation				Causal Structure Learning					Evaluation for Effect Estimation	
	Tool	CausalML	EconML	DoWhy	CauseBox	CausalNex	pcalg	bnlearn	TETRAD	CausalDiscovery	Causality-Benchmark	JustCause
Data	i.i.d	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
	IV	✓	✓	✓			✓	✓	✓			
	Networked											
	Time Series									✓		
Methods	Propensity Score			✓	✓							✓
	Tree-based	✓	✓	✓	✓			✓				
	Meta-Learner	✓	✓	✓								✓
	Doubly ML		✓	✓								
	Doubly Robust		✓	✓			✓					✓
	IV	✓	✓	✓								
	Mediation											
	Graph					✓	✓	✓	✓	✓		
	Pairwise				✓					✓		
Metrics	PEHE		User-Input Metrics			✓						✓
	RMSE	✓			✓	✓		✓		✓	✓	✓
	MAE	✓			✓	✓						✓
	Bias				✓						✓	✓
	Coverage										✓	
	Confidence Interval										✓	
	Aggregating Score										✓	
	Refutation											
	SID									✓		
	SHD						✓	✓	✓	✓		
	Classification					✓	✓	✓	✓	✓		

10) *TETRAD*[12] [143] is a drag and drop suite to perform causal structure learning. It can take datasets containing both continuous and categorical variables, including time-series data. It supports algorithms to search for structural relations. It also supports measuring unmeasured confounders, simulating data from a statistical model, predicting the effects on other variables of interventions or perturbations on one or more variables, and computing the probability distribution of any variable conditional on specified values of any other set of variables. It is developed in Java.

11) *pcalg*[13] [144] is a toolbox developed in R for causal structure learning and causal effect estimation using graphical models. For structural learning, pcalg supports multiple algorithms, including PC, FCI, RFCI, and GIES, whereas for causal effect estimation, it includes the IDA algorithm, the generalized backdoor criterion, and the generalized adjustment criterion.

Comparisons of these tools w.r.t. data, supported methods, and metrics can be found in Table III. An important notion is that data types supported by most of these tools are limited to i.i.d. data and data with IV. Furthermore, current evaluation-oriented tools do not support metrics for heterogeneous effect estimation and for causal structure learning tasks. In addition, none of these tools are designed to evaluate causality-aware ML tasks, e.g., causal interpretability.

VI. OPEN PROBLEMS AND CHALLENGES

Still an early-stage research in AI, CL presents both great opportunities and multifaceted challenges. We list a few to bring to the front some major concerns in the field.

A. Evaluation of Intermediate Steps

The importance of validating intermediate steps in a CL pipeline is often overlooked. For example, it is critical to precisely infer propensity scores as it is an intermediate step in many popular causal methods, e.g., propensity score matching. Therefore, an evaluation pipeline that effectively evaluates intermediate steps is desired. In text analysis, as text is interpretable and can be understood by humans, interpretable balance metrics, such as standardized difference in means and/or human judgments of propensity scores, can be used to assess the predicted propensity scores or matching results [145]. How to improve and standardize human judgment experiments, however, remains an open problem.

B. Evaluation for Networked Data

Networked data (e.g., social network) have been of great interest to researchers due to their ubiquity in real world. Some representative tasks include community detection and predicting relationships between individuals in the network. Despite the growing interest in identifying and estimating causal effects in networked data, methodologies and their evaluation have not kept apace [146]. Due to the interference and peer effects between units, standard structural causal models are not equipped to deal with dependence across individuals and describe the data generating process of networked data. This further impedes the research in causal structure learning for networked data. On the positive side, many evaluation metrics for causal inference and CL with i.i.d. data can still be used to benchmark methods for the corresponding tasks with networked data. The major challenge lies in the lack-of real-world datasets. Most of the existing work (see, e.g., [21] and [147]) relies on synthetic data generated from graph models or semisynthetic data with predefined structural equations.

C. Constructed Observational Studies

Constructed data—data gathered from both randomized and non-randomized experiments with similar subjects and

---

[12][Online]. Available: http://www.phil.cmu.edu/tetrad/
[13][Online]. Available: https://cran.r-project.org/web/packages/pcalg/index.html

settings—have been used in many areas, such as economics [3], marketing [148], and education [149]. AI researchers have also started using constructed data, as evidenced by recent works that leverage the advantages of both RCTs and observational studies to enhance online A/B tests (see, e.g., [150]). Using constructed data for CL makes up for the limitations in using only observational data or RCTs and helps make better decisions. Furthermore, it reveals possible solutions to evaluate intermediate steps and validate the process of model selection and hyperparameter tuning. It also allows us to evaluate causal effect estimation (e.g., ITE) when the ground truth is not available.

D. Evaluation of Long-Term Effects

Short- and long-term effects can differ in a number of ways. For example, user behavior changes as they learn and adapt to new environment. In a social network, user behavior is easily influenced by people in their network though the change may take time to reach its full effect [151]. Different from evaluating short-term effects, evaluating long-term effects is more cumbersome as we need to consider the factor of time. Current evaluation methods directly adopt pipelines in short-term effect estimation, e.g., MAE of the estimated effect at a specific time step. We need metrics tailored to the unique challenges in the long-term effect estimation. For example, when is the best target inferential point to estimate the effect? The answer reveals how long it takes a user to discover a new feature in a recommender system and if in-product education is needed to expedite the uptake. Many issues may arise given the nature of the data available to investigate long-term effects [152]. For instance, it is possible that a treatment effect depends on the previous levels of covariates; therefore, evaluation measures that can test for the presence and strength of effect modification are preferred. Sensitivity analysis and model validation are especially important in this task due to the issues of misspecifying the direction of causal pathways and censoring, e.g., people lost to follow-up over time.

E. Model Validation

Causal inference is generally an unsupervised learning technique, which makes the model validation challenging. With the required (and often unverifiable) causal assumptions, a valid and unbiased causal estimate can only be assured after the sensitivity and robustness of an estimate against these assumptions are checked. For example, despite the promising research progress on addressing unobserved confounding, current approaches are less readily available (or impossible) to remove all confounding. Even for unbiased estimators in the presence of unobserved confounders, assumptions required for causal identification and estimation can be violated. Therefore, being able to determine the robustness of causal methods is crucial for theory testing and model development. In addition, knowing to what extent the conclusions drawn from those causal studies are sensitive to potential confounding and violations of assumptions can help with policy decision making [153]. Although there exists an extensive literature of sensitivity analysis [154], most results are limited to specific model structures and on a case-by-case basis, whereas a general-purpose algorithmic framework for

sensitivity analysis is still lacking to account for the ubiquity of causal questions in the sciences and AI [155].

F. CL Evaluation Tools

The number of metrics and datasets for CL is fast growing. A pressing need for an open-source platform is thereby raised to promote objective, transparent, and independent evaluation of algorithms and to support broad collaborations. Such a platform is expected to include the ever-growing evaluation metrics, procedures, and datasets in causal effect estimation, causal structure learning, and causality-aware ML tasks. Model validation tools such as sensemakr[14] are also in need to better understand the robustness of a causal model. By providing convenient access to the benchmarking repository for evaluating CL algorithms, we can enlarge the collaboration in the community for the development of new metrics and datasets.

G. Evaluation of Causal Interpretability

Evaluating interpretable models is particularly difficult since interpretability does not have a unified definition [26]. Therefore, no specific measure can be used to fully assess models' interpretability from various aspects. It is even more challenging to evaluate models for causal interpretability due to the lack of ground-truth causal relations between the components of the model or the causal effect of one component on another. Therefore, crafting new datasets with such ground truth is a vital task for the field to move forward.

VII. Conclusion

CL is essential for AI to uncover causal mechanisms underlying real-world problems and to achieve human-level intelligence. Still at its infancy stage, research in CL is facing many obstacles. For example, CL has burgeoned, so far, without a proper benchmarking pipeline to support fair and transparent evaluations of emerging research contributions. To bridge this gap, in this survey, we provide a comprehensive review of the existing methods for evaluating the fundamental tasks in causal inference and causality-aware ML tasks and discuss potential limitations. We follow the evaluation pipeline in conventional ML to focus on the widely used datasets, evaluation metrics, protocols, and causal inference tools/packages. We conclude the survey with prominent open problems and challenges that await future research. To this end, we hope to broaden the discussions about open-source platforms/software that aim to promote CL research via fair, transparent, independent, and easy-to-use evaluation procedures.

We plan to develop a Causal Inference Evaluation Toolbox that consists of mainstream CL algorithms such as counterfactual regression (CFR) [31], causal effect variational autoencoder [39], Bayesian regression trees [156], causal forest [141], perfect match [48], disentangled representations for CFR [157], and similarity preserved IET [158]. This toolbox will showcase results of previously mentioned evaluation metrics for multiple benchmark datasets such as Jobs, IHDP, and News dataset.

---

[14][Online]. Available: https://cran.r-project.org/web/packages/sensemakr/index.html

REFERENCES

[1] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] L. Cheng, R. Guo, and H. Liu, "Robust cyberbullying detection with causal interpretation," in *Proc. Companion Proc. World Wide Web Conf.*, 2019, pp. 169–175.

[3] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *Amer. Econ. Rev.*, vol. 76, pp. 604–620, 1986.

[4] R. H. Dehejia and S. Wahba, "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *J. Amer. Statist. Assoc.*, vol. 94, no. 448, pp. 1053–1062, 1999.

[5] D. Heckerman, C. Meek, and G. Cooper, "A Bayesian approach to causal discovery," in *Innovations in Machine Learning*. New York, NY, USA: Springer, 2006, pp. 1–28.

[6] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *J. Comput. Graph. Statist.*, vol. 20, no. 1, pp. 217–240, 2011.

[7] S. Mani and G. F. Cooper, "Causal discovery from medical textual data," in *Proc. AMIA Symp.*, 2000, pp. 542–546.

[8] Cross-Disorder Group of the Psychiatric Genomics Consortium, "Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis," *Lancet*, vol. 381, no. 9875, pp. 1371–1379, 2013.

[9] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," *Rev. Econ. Statist.*, vol. 86, no. 1, pp. 4–29, 2004.

[10] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," *Epidemiology*, vol. 11, pp. 550–560, 2000.

[11] H. Ma and R. Jm, *Causal Inference*. Boca Raton, FL, USA: CRC Press, 2018.

[12] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *J. Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[13] J. Li, O. R. Zaïane, and A. Osornio-Vargas, "Discovering statistically significant co-location rules in datasets with extended spatial objects," in *Proc. Int. Conf. Data Warehousing Knowl. Discovery*, 2014, pp. 124–135.

[14] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," 2018, *arXiv:1801.04016*.

[15] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *J. Amer. Statist. Assoc.*, vol. 100, no. 469, pp. 322–331, 2005.

[16] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 1–37, 2021.

[17] N. Kallus and A. Zhou, "Confounding-robust policy improvement," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9289–9299.

[18] L. Cheng, R. Moraffah, R. Guo, K. Candan, A. Raglin, and H. Liu, "A practical data repository for causal learning with big data," in *Proc. BenchCouncil Int. Symp. Benchmarking, Meas. Optim.*, 2019, pp. 234–248.

[19] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[20] B. Schölkopf, "Causality for machine learning," 2019, *arXiv:1911.10500*.

[21] R. Guo, J. Li, and H. Liu, "Learning individual causal effects from networked observational data," in *Proc. Int. Conf. Web Search Data Mining*, 2020, pp. 232–240.

[22] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *Trans. Knowl. Discov. Data*, vol. 15, no. 5, pp. 1–46, 2021.

[23] P. Spirtes and K. Zhang, "Causal discovery and inference: Concepts and recent methodological advances," *Appl. Informat.*, vol. 3, no. 1, 2016, Art. no. 3.

[24] R. Moraffah et al., "Causal inference for time series analysis: Problems, methods and evaluation," *Knowl. Inf. Syst.*, vol. 63, pp. 3041–3085, 2021.

[25] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," 2020, *arXiv:2010.03240*.

[26] R. Moraffah, M. Karami, R. Guo, A. Ragliny, and H. Liu, "Causal interpretability for machine learning—Problems, methods and evaluation," *ACM SIGKDD Explorations Newslett.*, vol. 22, pp. 18–33, 2020.

[27] Y. Shimoni, C. Yanover, E. Karavani, and Y. Goldschmnidt, "Benchmarking framework for performance-evaluation of causal inference analysis," 2018, *arXiv:1802.05046*.

[28] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[29] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1414–1423.

[30] A. M. Puli and R. Ranganath, "Generalized control functions via variational decoupling," 2019, *arXiv:1907.03451*.

[31] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3076–3085.

[32] V. Rakesh, R. Guo, R. Moraffah, N. Agarwal, and H. Liu, "Linked causal variational autoencoder for inferring paired spillover effects," in *Proc. Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1679–1682.

[33] Y. Zhao, X. Fang, and D. Simchi-Levi, "Uplift modeling with multiple treatments and general response types," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 588–596.

[34] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modeling: A review of the literature," in *Proc. Int. Conf. Predictive Appl. APIs*, 2017, pp. 1–13.

[35] N. J. Radcliffe, "Using control groups to target on predicted lift: Building and assessing uplift models," *Direct Marketing Analytics J.*, vol. 1, 2007, Art. no. 1421.

[36] D. Card and A. B. Krueger, "Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania," *Amer. Econ. Rev.*, vol. 84, no. 4, pp. 772–793, 1994.

[37] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: Methods and benchmarks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1103–1204, 2016.

[38] A. Dixit et al., "Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens," *Cell*, vol. 167, no. 7, pp. 1853–1866, 2016.

[39] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6446–6456.

[40] N. Kallus, X. Mao, and M. Udell, "Causal inference with noisy and missing covariates via matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6921–6932.

[41] A. M. Gentzel, P. Pruthi, and D. Jensen, "How and why to use experimental data to evaluate methods for observational causal inference," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3660–3671.

[42] J. A. Smith and P. E. Todd, "Does matching overcome LaLonde's critique of nonexperimental estimators?," *J. Econometrics*, vol. 125, pp. 305–353, 2005.

[43] P. R. Hahn, V. Dorie, and J. S. Murray, "Atlantic causal inference conference (ACIC) data analysis challenge 2017," 2019, *arXiv:1905.09515*.

[44] G. J. Duncan, J. Brooks-Gunn, and P. K. Klebanov, "Economic deprivation and early childhood development," *Child Develop.*, vol. 65, no. 2, pp. 296–318, 1994.

[45] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *Quart. J. Econ.*, vol. 120, no. 3, pp. 1031–1083, 2005.

[46] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 785–794.

[47] J. Yoon, J. Jordon, and M. van der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[48] P. Schwab, L. Linhardt, and W. Karlen, "Perfect match: A simple method for learning representations for counterfactual inference with neural networks," 2018, *arXiv:1810.00656*.

[49] D. Galagate, J. Schafer, and M. D. Galagate, "Package 'causaldrf'," 2015.

[50] J. D. Angrist and J.-S. Pischke, *Mastering Metrics: The Path From Cause to Effect*. Princeton, NJ, USA: Princeton Univ. Press, 2014.

[51] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *J. Amer. Statist. Assoc.*, vol. 91, no. 434, pp. 444–455, 1996.

[52] D. T. Campbell, "Reforms as experiments," *Amer. Psychol.*, vol. 24, no. 4, pp. 409–429, 1969.

[53] J. D. Angrist and A. B. Krueger, "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples," *J. Amer. Statist. Assoc.*, vol. 87, no. 418, pp. 328–336, 1992.

[54] J. D. Angrist and G. W. Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *J. Amer. Statist. Assoc.*, vol. 90, no. 430, pp. 431–442, 1995.

[55] N. Nunn and L. Wantchekon, "The slave trade and the origins of mistrust in Africa," *Amer. Econ. Rev.*, vol. 101, no. 7, pp. 3221–52, 2011.

[56] C. Carpenter and C. Dobkin, "The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age," *Amer. Econ. J. Appl. Econ.*, vol. 1, no. 1, pp. 164–82, 2009.

[57] A. C. Eggers, R. Freier, V. Grembi, and T. Nannicini, "Regression discontinuity designs based on population thresholds: Pitfalls and solutions," *Amer. J. Political Sci.*, vol. 62, no. 1, pp. 210–229, 2018.

[58] R. Jacob, P. Zhu, M.-A. Somers, and H. Bloom, *A Practical Guide to Regression Discontinuity*. New York, NY, USA: MDRC, 2012.

[59] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[60] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott, "Inferring causal impact using Bayesian structural time-series models," *Ann. Appl. Statist.*, vol. 9, no. 1, pp. 247–274, 2015.

[61] M. Auffhammer and R. Kellogg, "Clearing the air? The effects of gasoline content regulation on air quality," *Amer. Econ. Rev.*, vol. 101, no. 6, pp. 2687–2722, 2011.

[62] Y. Ma and V. Tresp, "Causal inference under networked interference and intervention policy enhancement," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3700–3708.

[63] K. Chantala and J. Tabor, "National longitudinal study of adolescent health: Strategies to perform a design-based analysis using the add health data," 1999.

[64] L. Takac and M. Zabovsky, "Data analysis in public social networks," in *Proc. Int. Sci. Conf. Int. Workshop Present Day Trends Innov.*, 2012, pp. 1–6.

[65] R. Guo, J. Li, and H. Liu, "Counterfactual evaluation of treatment assignment functions with networked observational data," in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 271–279.

[66] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible AI algorithms: Issues, purposes, and challenges," *J. Artif. Intell. Res.*, vol. 71, pp. 1137–1181, 2021.

[67] A. Volfovsky, E. M. Airoldi, and D. B. Rubin, "Causal inference for ordinal outcomes," 2015, *arXiv:1501.01234*.

[68] L. Cheng, R. Guo, and H. Liu, "Long-term effect estimation with surrogate representation," in *Proc. Int. Conf. Web Search Data Mining*, 2021, pp. 274–282.

[69] P. Spirtes *et al.*, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.

[70] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 459–466.

[71] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[72] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[73] M. de Jongh and M. J. Druzdzel, "A comparison of structural distance measures for causal Bayesian network models," in *Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science Series*. Warsaw, Poland: Academic, 2009, pp. 443–456.

[74] J. Peters and P. Bühlmann, "Structural intervention distance for evaluating causal graphs," *Neural Comput.*, vol. 27, no. 3, pp. 771–799, 2015.

[75] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, 2006.

[76] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: Identification and confidence intervals," *J. Roy. Statist. Soc. Ser. B. Statist. Methodol.*, vol. 78, no. 5, pp. 947–1012, 2016.

[77] S. Shimizu *et al.*, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *J. Mach. Learn. Res.*, vol. 12, pp. 1225–1248, 2011.

[78] D. Bacciu, T. A. Etchells, P. J. Lisboa, and J. Whittaker, "Efficient identification of independence networks using mutual information," *Comput. Statist.*, vol. 28, no. 2, pp. 621–646, 2013.

[79] A. Gentzel, D. Garant, and D. Jensen, "The case for evaluating causal models using interventional measures and empirical data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019.

[80] M. Peyrard and R. West, "A ladder of causal distances," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 2012–2018.

[81] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[82] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[83] J. Mitrovic, D. Sejdinovic, and Y. W. Teh, "Causal inference via kernel deviance measures," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 6986–6994.

[84] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[85] M. Kocaoglu, A. Dimakis, and S. Vishwanath, "Cost-optimal learning of causal graphs," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1875–1884.

[86] K. Shanmugam, M. Kocaoglu, A. G. Dimakis, and S. Vishwanath, "Learning causal graphs with small interventions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 3195–3203.

[87] O. Ahmed *et al.*, "CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[88] T. Van den Bulcke *et al.*, "SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 43.

[89] J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz-Marí, and G. Camps-Valls, "The causality for climate competition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 110–120.

[90] U. Schaechtle, K. Stathis, and S. Bromuri, "Multi-dimensional causal discovery," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1649–1655.

[91] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao, "Causal discovery from temporally aggregated time series," *Uncertain Artif. Intell.*, vol. 2017, 2017, Art. no. 269.

[92] Z. Pan, Y. Liang, J. Zhang, X. Yi, Y. Yu, and Y. Zheng, "HyperST-Net: Hypernetworks for spatio-temporal forecasting," 2018, *arXiv:1809.10889*.

[93] D. Entner and P. O. Hoyer, "On causal discovery from time series data using FCI," in *Proc. 5th Eur. Workshop Probab. Graph. Models*, 2010, pp. 121–128.

[94] D. Malinsky and D. Danks, "Causal discovery algorithms: A practical guide," *Philosophy Compass*, vol. 13, no. 1, 2018, Art. no. e 12470.

[95] F. Lattimore, T. Lattimore, and M. D. Reid, "Causal bandits: Learning good interventions via causal inference," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1181–1189.

[96] S. Lee and E. Bareinboim, "Structural causal bandits: Where to intervene?," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2568–2578.

[97] L. Cheng, A. Mosallanezhad, P. Sheth, and H. Liu, "Causal learning for socially responsible AI," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 4374–4381.

[98] Z. C. Lipton, "The Mythos of model interpretability," Queue, vol. 16, no. 3, pp. 31–57, 2018.

[99] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[100] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2376–2384.

[101] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (CACE)," 2019, *arXiv:1907.07165*.

[102] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," 2018, *arXiv:1811.04376*.

[103] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," 2018, *arXiv:1802.00541*.

[104] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[105] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4066–4076.

[106] Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-fairness: A unified framework for measuring causality-based fairness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 3399–3409.

[107] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Databases*, 2021, pp. 650–665.

[108] R. Mc Grath *et al.*, "Interpretable credit application predictions with counterfactual explanations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018.

[109] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 607–617.

[110] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, "Multi-modal explanations by predicting counterfactuality in videos," 2018, *arXiv:1812.01263*.

[111] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Survey on causal-based machine learning fairness notions," 2020, *arXiv:2010.09553*.

[112] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *Proc. Int. World Wide Web Conf.*, 2019, pp. 2907–2914.

[113] J. Zhang and E. Bareinboim, "Fairness in decision-making—The causal explanation formula," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2037–2045.

[114] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (IM) possibility of fairness," 2016, *arXiv:1609.07236*.

[115] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, vol. 96, 1996, pp. 202–207.

[116] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, pp. 841–887, 2018.

[117] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[118] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database," Jan. 2020. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[119] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[120] K. Lang, "20 newsgroups," 2008. [Online]. Available: http://qwone.com/jason/20Newsgroups/

[121] IMDb, "IMDb dataset," 2020. [Online]. Available: https://www.imdb.com/interfaces/

[122] AWS, "Amazon customer reviews dataset," 2020. [Online]. Available: https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[123] UCI, "UCI machine learning repository," 2020. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[124] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[125] A. Flores, K. Bechtel, and C. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals and it's biased against blacks"," *Federal Probation*, vol. 80, pp. 38–46, Sep. 2016.

[126] R. Guo, X. Zhao, A. Henderson, L. Hong, and H. Liu, "Debiasing grid-based product search in E-commerce," in *Proc. 26th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2020, pp. 2852–2860.

[127] Z. Wang, X. Yin, T. Li, and L. Hong, "Causal meta-mediation analysis: Inferring dose-response function from summary statistics of many randomized experiments," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2625–2635.

[128] X. Wu, H. Chen, J. Zhao, L. He, D. Yin, and Y. Chang, "Unbiased learning to rank in feeds recommendation," in *Proc. Int. Conf. Web Search Data Mining*, 2021, pp. 490–498.

[129] T. Joachims, A. Swaminathan, and M. De Rijke, "Deep learning with logged bandit feedback," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[130] Z. Hu, Y. Wang, Q. Peng, and H. Li, "Unbiased LambdaMART: An unbiased pairwise learning-to-rank algorithm," in *Proc. World Wide Web Conf.*, 2019, pp. 2830–2836.

[131] Q. Ai, K. Bi, C. Luo, J. Guo, and W. B. Croft, "Unbiased learning to rank with unbiased propensity estimation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 385–394.

[132] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[133] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," in *Proc. Int. Florida Artif. Intell. Res. Soc. Conf.*, 2019, pp. 413–418.

[134] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proc. Int. Conf. Web Search Data Mining*, 2008, pp. 87–94.

[135] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview," in *Proc. Learn. Rank Challenge*, 2011, pp. 1–24.

[136] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "CausaLML: Python package for causal machine learning," 2020, *arXiv:2002.11631*.

[137] M. Research, "EconML: A python package for ML-based heterogeneous treatment effects estimation," 2019. [Online]. Available: https://github.com/microsoft/EconML

[138] "DoWhy: A Python package for causal inference," 2019. [Online]. Available: https://github.com/microsoft/dowhy

[139] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9472–9483.

[140] D. Kalainathan and O. Goudet, "Causal discovery toolbox: Uncover causal relationships in python," 2019, *arXiv:1903.02278*.

[141] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Amer. Statist. Assoc.*, vol. 113, pp. 1228–1242, 2018.

[142] M. Scutari, "Learning Bayesian networks with the bnlearn R package," *J. Statist. Softw.*, vol. 35, pp. 1–22, 2010.

[143] J. D. Ramsey *et al.*, "TETRAD—A toolbox for causal discovery," in *Proc. 8th Int. Workshop Climate Informat.*, 2018, pp. 89–92.

[144] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, "Causal inference using graphical models with the R package pcalg," *J. Statist. Softw.*, vol. 47, no. 11, pp. 1–26, 2012.

[145] K. Keith, D. Jensen, and B. O'Connor, "Text and causal inference: A review of using text to remove confounding from causal estimates," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5332–5344.

[146] E. L. Ogburn, O. Sofrygin, I. Díaz, and M. J. Van der Laan, "Causal inference for social network data," in *Advances In Water Resources*. Southampton, U.K.: Computational Mechanics Publications, 2017.

[147] B. Bevilacqua, Y. Zhou, and B. Ribeiro, "Size-invariant graph representations for graph classification extrapolations," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 837–851.

[148] B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky, "A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook," *Marketing Sci.*, vol. 38, no. 2, pp. 193–225, 2019.

[149] W. R. Shadish, M. H. Clark, and P. M. Steiner, "Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments," *J. Amer. Statist. Assoc.*, vol. 103, no. 484, pp. 1334–1344, 2008.

[150] L. Ye, Y. Lin, H. Xie, and J. Lui, "Combining offline causal inference and online bandit learning for data driven decisions," 2020, *arXiv:2001.05699*.

[151] R. Kohavi and R. Longbotham, "Online controlled experiments and A/B testing," *Encyclopedia Mach. Learn. Data Mining*, vol. 7, no. 8, pp. 922–929, 2017.

[152] S. J. Newsome, R. H. Keogh, and R. M. Daniel, "Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty," *Statist. Med.*, vol. 37, no. 15, pp. 2367–2390, 2018.

[153] W. Liu, S. J. Kuramoto, and E. A. Stuart, "An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research," *Prevention Sci.*, vol. 14, no. 6, pp. 570–580, 2013.

[154] J. Huang, Nov. 2020. [Online]. Available: https://medium.com/data-science-at-microsoft/causal-inference-part-3-of-3-model-validation-and-applications-c84764156a29

[155] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim, "Sensitivity analysis of linear structural causal models," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, Jun. 2019, pp. 1252–1261.

[156] P. R. Hahn, J. S. Murray, and C. M. Carvalho, "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)," *Bayesian Anal.*, vol. 15, no. 3, pp. 965–1056, 2020.

[157] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[158] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 2633–2643.

Lu Cheng received the B.Eng. degree in logistic and systems engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2015, and the M.Eng. degree in industrial engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2017. She is currently working toward the Ph.D. degree in computer science and engineering with Arizona State University, Tempe, AZ, USA.

She has authored or coauthored research papers in premier conferences of artificial intelligence (AI), data mining, and machine learning. Her research interests include socially responsible AI, causal learning, and data mining.

Ms. Cheng is a Student Member of the Association for Computing Machinery.

Ruocheng Guo received the Ph.D. degree in computer engineering from Arizona State University, Tempe, AZ, USA, in 2021.

He is currently an Assistant Professor of Data Science with the City University of Hong Kong, Hong Kong. His research interests include causal machine learning toward fair, interpretable and generalizable artificial intelligence, causal inference, and data mining.

Dr. Guo is a Member of the Association for Computing Machinery, the Society for Industrial and Applied Mathematics, and the Association for the Advancement of Artificial Intelligence.

Raha Moraffah received the B.S. degree in computer science and engineering from the Sharif University of Technology, Tehran, Iran, in 2015. She is currently working toward the Ph.D. degree in computer science and engineering with Arizona State University, Tempe, AZ, USA.

Her research interests include causal inference, causal machine learning, and adversarial learning.

Ms. Moraffah is a Student Member of the Association for Computing Machinery.

Paras Sheth received the B.Tech. degree in computer science and engineering from the Institute of Engineering and Management, Kolkata, India, in 2017. He is currently working toward the Ph.D. degree in computer science and engineering with Arizona State University, Tempe, AZ, USA.

His research interests include causal inference, data mining, and causal machine learning.

K. Selçuk Candan received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, in 1997.

He is currently a Professor of Computer Science and Engineering and the Director of the Center for Assured and Scalable Data Engineering, Arizona State University, Tempe, AZ, USA. His research interests include data management, analysis, and scalable machine learning.

Dr. Candan is a Distinguished Scientist of the Association for Computing Machinery.

Huan Liu (Fellow, IEEE) received the B.Eng. degree in computer science and electrical engineering from Shanghai Jiaotong University, Shanghai, China, in 1983, and the Ph.D. degree in computer science from the University of Southern California, Los Angeles, CA, USA, in 1989.

He is currently a Professor of Computer Science and Engineering with Arizona State University, Tempe, AZ, USA. His well-cited publications include books, book chapters, encyclopedia entries, and conference and journal papers. His research interests include data mining, machine learning, social computing, and artificial intelligence, investigating problems that arise in many real-world applications with high-dimensional data of disparate forms.

Dr. Liu is a Fellow of the Association for Computing Machinery, the Association for the Advancement of Artificial Intelligence, and the American Association for the Advancement of Science.