Received: 26 July 2021 | Revised: 7 December 2021 | Accepted: 22 December 2021

DOI: 10.1002/widm.1449

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

WILEY

Methods and tools for causal discovery and causal inference

Ana Rita Nogueira[1,2] | Andrea Pugnana[3] | Salvatore Ruggieri[4] | Dino Pedreschi[4] | João Gama[1]

[1]LIAAD, INESC TEC, Porto, Portugal

[2]PDCC, Faculdade de Ciências da Universidade do Porto, Porto, Portugal

[3]Scuola Normale Superiore, Pisa, Italy

[4]Università di Pisa, Pisa, Italy

Correspondence
Ana Rita Nogueira, LIAAD, INESC TEC, Porto, 4200465, Portugal.
Email: ana.r.nogueira@inesctec.pt

Funding information
This work was partially supported by the European Commission funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437). This work was also partially supported by the Fundação para a Ciência e Tecnologia, Grant/Award Number: SFRH/BD/146197/20198.

Edited by: Sushmita Mitra, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract

Causality is a complex concept, which roots its developments across several fields, such as statistics, economics, epidemiology, computer science, and philosophy. In recent years, the study of causal relationships has become a crucial part of the Artificial Intelligence community, as causality can be a key tool for overcoming some limitations of correlation-based Machine Learning systems. Causality research can generally be divided into two main branches, that is, causal discovery and causal inference. The former focuses on obtaining causal knowledge directly from observational data. The latter aims to estimate the impact deriving from a change of a certain variable over an outcome of interest. This article aims at covering several methodologies that have been developed for both tasks. This survey does not only focus on theoretical aspects. But also provides a practical toolkit for interested researchers and practitioners, including software, datasets, and running examples.

This article is categorized under:

Algorithmic Development > Causality Discovery
Fundamental Concepts of Data and Knowledge > Explainable AI
Technologies > Machine Learning

KEYWORDS

causal discovery, causal inference, causality

1 | INTRODUCTION

From the very beginning, the study of nature drove human knowledge. In particular, the search for causes of natural phenomena animated several philosophers in Ancient Greece, such as Plato or Aristotle. For instance, in Plato (1961), the so-called "inquiry into Nature" consisted in a quest for "the causes of each thing; why each thing comes into existence, why it goes out of existence, why it exists". Similarly, nowadays, the study of causality covers different disciplines, aiming to answer "Why?" questions. A well-known mantra in these fields is that correlation does not imply causation. In general, assuming that if there is a correlation, there is also causation is a fallacy due to omitted data or links for which (biased) human reasoning leads to erroneous conclusions. The same goes for the opposite relation:

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama contributed equally to this study.

causation does not imply correlation. There are cases in which, although there is a strong causal relationship between events, there is no evidence of correlation in a specific sample. There are many examples of causal fallacies; a notorious one dates back to Ancient Greece, that is, the Aristotelian theory of spontaneous generation (Comandé, 2018). According to this theory, some organisms could be generated from inanimate matter. Despite its wrong conclusions, this thesis was deeply empirical, as it was developed after observing that flies appeared in the presence of rotting meat. Only, in the 19th century, this theory was falsified by Pasteur. Similarly, modern data mining techniques could be prone to the same fallacies as they are built over correlations among variables.

The fact that artificial intelligence (AI) systems are based on simple associations is problematic, as they are becoming ubiquitous in daily human activities. As a consequence, an urge for trustworthy machine learning (ML) tools arose. Due to this, the scientific community has made a considerable effort to study causality, shifting the focus from philosophy and empirical experiments to AI and ML domains. As a fact, causality is a potential tool to solve some of the current ML limitations (Pearl, 2018).

However, given its inherent multi-disciplinary history, the study of causality is fragmented. Several contributions came from different fields, such as epidemiology, economics, statistics, computer science, and so forth. Still, two main tasks can be distinguished within the realm of causality: causal discovery and causal inference. Starting from a set of observational data, the former tries to infer the causal relationship across the different variables in the dataset. The latter focuses on testing whether two variables are related and assessing the impact of one on the other. Clearly, these two tasks are antipodal: on the one hand, causal discovery does not assume any relationship among involved variables; rather, it is inferred directly from a set of data. On the other hand, causal inference assumes a relationship among variables and tries to test and quantify the actual relationship in the available data.

Given that the study of causality has always been of interest for many scientific fields, a few attempts to review the state of the art in causality have been proposed in the past, each of them focusing on specific aspects or application areas. Table 1 provides a summary of existing surveys and their topics. We distinguish theory, datasets, tools, metrics, and examples for both causal discovery and causal inference.

In light of this, the present work has the ambitious goal of covering all of the aspects (theory, datasets, software tools, evaluation metrics, and running examples) for the main techniques of causal discovery and causal inference. As such, it should be an initial reference for both researchers and practitioners interested in the main pointers to methodologies, data, and tools for causal discovery and causal inference. A companion website[1] supplements this survey with an updated list of datasets, tools, and practical examples (with scripts in Python and R).

The paper is structured as follows. In Section 2, some basic definitions and notations are introduced. In Section 3, causal discovery techniques, tools, datasets, metrics, and examples are presented, organized by data type (cross-sectional, time-series, longitudinal). Section 4 covers causal inference techniques for several causal effects, tools, datasets, and a running example. Some remarks regarding the intersection between ML and causality are presented in Section 5, where some of the current open issues are also highlighted. Finally, conclusions are drawn.

2 | DEFINITIONS AND NOTATIONS

Causality can be defined as the influence by which an event contributes to the production of other events. The cause is responsible for creating the effect, and the effect is a consequence of the cause taking place. For instance, if we consider two different events A and B as an example, where the latter is a consequence of the former, A is a necessary requirement for B to exist, but B is not required for A to happen. Therefore, studying causality means understanding how different events, involving different variables, are related among themselves.

It is important to note that causality is a rather broad concept and it covers different fields. It combines statistics, machine learning, data mining, and several other quantitative disciplines to search for potential cause–effect relationships in observational data (Guo et al., 2020). As previously described, it is typically divided into causal discovery and causal inference. Causal discovery is responsible for analyzing and creating models that illustrate the relationships inherent in the data. Causal inference aims to study the possible effects of altering a given system (Yao et al., 2021).

Generally speaking, causal models are "mathematical models representing causal relationships within an individual system or population" (Hitchcock, 2020). Causal relationships entail the probabilistic (in)dependence of variables, and the effects of interventions (change on some variables) or hypothetical interventions, such as counterfactual claims.

We introduce in this section some of the main causal models, starting from the necessary background on graphs, as they are a powerful tool to represent visually the relationships across variables in a system. The following definitions

TABLE 1 Overview of survey papers on causality

	Causal discovery					Causal inference				
Paper title	Theory	Data	Software	Evaluation metrics	Practical examples	Theory	Data	Software	Evaluation metrics	Practical examples
A Survey of Learning Causality with Data: Problems and Methods (Guo et al., 2020)	✓			✓	✓	✓				✓
Causal Interpretability for Machine Learning: Problems, Methods, and Evaluation (Moraffah et al., 2020)						✓	✓		✓	
Causal Inference for Time Series Analysis: Problems, Methods, and Evaluation (Moraffah et al., 2021)	✓	✓		✓		✓	✓		✓	
Review of Causal Discovery Methods Based on Graphical Models (Glymour et al., 2019)	✓				✓					
Causal discovery in machine learning: Theories and applications (Nogueira et al., 2021)	✓		✓	✓						
A Survey on Causal Inference (Yao et al., 2021)						✓	✓	✓		
Toward Causal Representation Learning (Schölkopf et al., 2021)	✓						✓			

are pretty self-explanatory and can be found in several books covering causality, such as Pearl (2009), Peters et al. (2017), and Spirtes et al. (2000).

A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is defined by a set of nodes (or vertices) V and a set of edges $\mathcal{E} \subseteq \{(\{U,V\},M) \mid U,V \in \mathbf{V}, U \neq V, M \in \mathbf{M}\}$, where $\mathbf{M}$ is a set of marks (or labels). In particular, an edge can be marked as directed, undirected, or bi-directed, respectively written as $U \rightarrow V$ or $V \rightarrow U$, $U - V$, and $U \leftrightarrow V$. Nodes U and V are said adjacent. The edge relation $\mathcal{E}$ is actually a partial function, that is, no more than one mark is assigned to adjacent nodes. A graph $\mathcal{G}$ is directed if all the edges are directed. It is a pattern if each edge is either directed or undirected. A node $U \in \mathbf{V}$ is a parent of another node $V \in \mathbf{V}$ if $U \rightarrow V \in \mathcal{E}$. The node V is said to be a child of node U. We write $\mathrm{Pa}(V)$ for the set of parents of V, and $\mathrm{Ch}(U)$ for the set of children of U. A (acyclic) path in $\mathcal{G}$ is a sequence of distinct vertices $V_1, \ldots, V_n$ such that an edge $(\{V_j, V_{j+1}\}, M_j)$ between two consecutive vertices is in $\mathcal{E}$, for $j = 1, \ldots, n-1$. When all the edges are directed as $V_j \rightarrow V_{j+1}$, the path is called a directed path. In such a case, the node V_1 is called an ancestor of V_n, while V_n is called a descendant of V_1. The set of all the ancestors of V is denoted as $\mathrm{An}(V)$ while the set of descendants is written as $\mathrm{De}(V)$. Notice that $V \in \mathrm{An}(V)$ and $V \in \mathrm{De}(V)$. A direct graph is called a directed acyclic graph (DAG) if there is no directed cycle, that is, there exists no pair of vertices $V \neq U$ with a directed path from V to U and from U to V.

DAGs were used by Pearl (1985) as a graphical representation for a constrained joint probability distribution of a collection of random variables. Let us consider p random variables $\mathbf{X} = (X_1, \ldots, X_p)$ with joint distribution $P(\mathbf{X})$. Let $P(X_i|\mathbf{S})$ be the marginal distribution of X_i conditional to $\mathbf{S} \subseteq \mathbf{X}$.

Definition 1. *Given a DAG $\mathcal{G} = (\boldsymbol{X}, \mathcal{E})$, the random variables $\boldsymbol{X}$ form a Bayesian network with respect to $\mathcal{G}$ if:*

$$P(\boldsymbol{X}) = \prod_{X \in \boldsymbol{X}} P(X|Pa(X)) \tag{1}$$

Bayesian networks are graphical representations of probabilistic relations among variables, where nodes represent the variables themselves, while edges represent the conditional dependencies among the involved variables. It is worth noticing that such a representation is convenient, as it allows to clearly model how the variables are related in the system. For example, let us consider the well-known example about kidney stones, where one is interested in understanding how drug usage influences the recovery of patients. Let T be a binary variable of treatment (treated/not treated), and Y the outcome (recovered/unrecovered), such that $T \rightarrow Y$. Let us also consider some confounding variable U, such that $U \rightarrow T$ and $U \rightarrow Y$. For instance, U could be the age of the patient or the size of the kidney stones or any other factor that might affect both the access to the treatment and the ability to recover. This set of information can be represented through the DAG shown in Figure 1.

The factorization formula (1) is equivalent for DAGs (Pearl, 1989) to the following Markov condition, stating that a variable is conditionally independent ($\perp\!\!\!\perp$) of each of its non-descendants, given its parents.

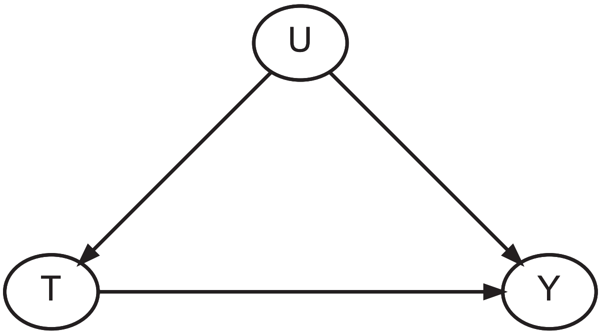


FIGURE 1 The kidney stones example represented as a DAG: T is the treatment indicator, Y is the outcome, and U is a confounder

Definition 2. (Markov Condition) *Given a DAG* $\mathcal{G}=(\boldsymbol{X},\mathcal{E})$, *the random variables* $\boldsymbol{X}$ *satisfy the Markov Condition if for every* $X\in\boldsymbol{X}$, $X \perp\!\!\!\perp \boldsymbol{X}\backslash(De(X)\cup Pa(X))\mid Pa(X)$.

The Markov Condition is not sufficient to read off arbitrary conditional (in)dependencies entailed by a Bayesian network. For this, we need d-separation. Let us first introduce the notion of blocking set.

Definition 3. (Blocking set) *A path* $V_1,\ldots,V_n$ *in a DAG* $\mathcal{G}$ *is blocked by a set of nodes* $\boldsymbol{Z}$ *(not containing neither* V_1 *nor* V_n*) if there exists a node* V_k *in the path such that one of the following conditions hold*:

(i) V_k *is a non-collider, that is,* $V_{k-1}\rightarrow V_k\rightarrow V_{k+1}$ *or* $V_{k-1}\leftarrow V_k\leftarrow V_{k+1}$ *or* $V_{k-1}\leftarrow V_k\rightarrow V_{k+1}$, *and* $V_k\in Z$;

(ii) V_k *is a collider, that is,* $V_{k-1}\rightarrow V_k\leftarrow V_{k+1}$, *and* $De(V_k)\cap\boldsymbol{Z}=\emptyset$, *that is, neither* V_k *nor any of its descendants is in* Z.

Definition 4. (d-separation) *In a DAG* $\mathcal{G}$, *we say that two sets of nodes* $\boldsymbol{L}$ *and* $\boldsymbol{M}$ *are d-separated by a third set of nodes* $\boldsymbol{Z}$, *where* $\boldsymbol{L}$, $\boldsymbol{M}$ *and* $\boldsymbol{Z}$ *are pairwise disjoint, if* $\boldsymbol{Z}$ *is blocking all the paths between nodes in* $\boldsymbol{L}$ *and* $\boldsymbol{M}$. *This is denoted as*: $\boldsymbol{L}\perp_{\mathcal{G}}\boldsymbol{M}\mid\boldsymbol{Z}$.

For example, the variable Y d-separates X and Z in the DAG of Figure 2. The factorization formula (1) is also equivalent for DAGs to the Global Markov Condition (Pearl, 1989),

Definition 5. (Global Markov Condition) *Given a DAG* $\mathcal{G}=(\boldsymbol{X},\mathcal{E})$, *the random variables* $\boldsymbol{X}$ *satisfy the Global Markov Condition if for every pairwise disjoint* $\boldsymbol{L},\boldsymbol{M},\boldsymbol{Z}\subseteq\boldsymbol{X}$, *if* $\boldsymbol{L}\perp_{\mathcal{G}}\boldsymbol{M}\mid\boldsymbol{Z}$ *then* $\boldsymbol{L}\perp\!\!\!\perp\boldsymbol{M}\mid\boldsymbol{Z}$.

The Faithfulness assumption reverses the direction of the above implication, so that conditionally independent variables are actually d-separated in the graph.

Definition 6. (Faithfulness) *Given a DAG* $\mathcal{G}=(\boldsymbol{X},\mathcal{E})$, *the random variables* $\boldsymbol{X}$ *satisfy the Faithfulness assumption if for every pairwise disjoint* $\boldsymbol{L},\boldsymbol{M},\boldsymbol{Z}\subseteq\boldsymbol{X}$, *if* $\boldsymbol{L}\perp\!\!\!\perp\boldsymbol{M}\mid\boldsymbol{Z}$ *then* $\boldsymbol{L}\perp_{\mathcal{G}}\boldsymbol{M}\mid\boldsymbol{Z}$.

Finally, the assumption of *Causal Sufficiency* states that all the common causes of a pair of nodes are measured. Although most methods rely on this assumption, it cannot always be satisfied. Due to this, some methods model the existence of latent variables.

Definition 7. (Causal Sufficiency) *For a pair of observed variables* X *and* Y, *all their common causes must also be observed in the data (and modeled in a graph* $\mathcal{G}$*).*

Edges in the Bayesian network model conditional probabilities (*condition on observing*), but they do not necessarily represent causal effects (*intervention*). Intuitively, a variable X has a causal effect on Y if manipulating X changes the distribution of Y. Causal Bayesian network extends the factorization formula (1) to account for *do*-interventions: $P(\mathbf{X}|do(\mathbf{W}=\mathbf{w}))=\prod_{X\in\mathbf{X}\backslash\mathbf{W}}P(X|\mathrm{Pa}(X))1_{\mathbf{W}=\mathbf{w}}$. The *do-operator*, denoted as $do(W=w)$ and introduced by Pearl (1995), represents the symbolic operation of setting the definition W to the constant value w (*atomic intervention*). Intervention distributions $P(X\mid do(W=w)$ are not necessarily equivalent to conditional distributions $P(X|W=w)$. Contrasted with this stochastic approach to causality (Pearl, 2009, sect. 1.4), there is the Laplacian's approach based on functional equations. We recall next the Structural Causal Models.

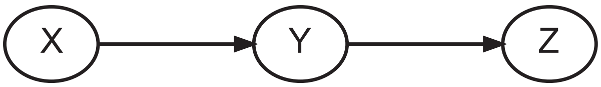


FIGURE 2 Example d-separation: Y d-separates X and Z

Definition 8. (Structural Causal Model) *Given a DAG $\mathcal{G} = (\boldsymbol{X}, \mathcal{E})$, a Structural Causal Model (SCM) defines the (endogenous) random variables* $\boldsymbol{X}$ *as functions of their parents*:

$$X_i := f_i(Pa(X_i), U_i), \quad i = 1, ..., p$$

and of (exogenous) independent random variables $U_1, ..., U_p$.

An assignment of values to the exogenous variables uniquely determines the values of all endogenous variables. Thus, a probability distribution P' over $U_1, ..., U_p$ induces a unique probability distribution P over X (*entailed distribution*).

An SCM is a Causal Bayesian network if interventions are modeled as follows (Peters et al., 2017).

Definition 9. (Intervention distribution) *The probability $P(\boldsymbol{X}|do(\boldsymbol{W} = \boldsymbol{w}))$ over a SCM is the distribution entailed by the SCM obtained by replacing, for $X_k = w_k$ in $\boldsymbol{W} = \boldsymbol{w}$, the definition $X_k := f_k(Pa(X_k), U_k)$ with $X_k := w_k$.*

Let us reconsider the kidney stones example. Performing a do-operations on the treatment variable consists of removing all the incoming links on the node T. Graphically, this is depicted in Figure 3. The causal effect of drug use can then be measured by comparing the intervention distributions $P(Y = 1|do(T = 1))$ and $P(Y = 1|do(T = 0))$.

Another well-known method for estimating causal effects is the potential outcome (PO) model, also known as Rubin causal model (RCM), as defined in Holland (1986). Introduced in Rubin (1974), it is very popular in economics and social sciences, and logically equivalent to SCM (Pearl, 2009). Unless stated otherwise, we restrict to a binary treatment. The PO model assumes random variables $\mathbf{X}, Y_0, Y_1, T$, where $\mathbf{X}$ are observed covariates about atomic research objects, called units or individuals; Y_1 is the outcome after being treated; Y_0 is the outcome after not being treated; and T is the actual intervention, with $T = 1$ for treated and $T = 0$ for not treated. The observed outcome is $Y = Y_T = Y_0 + T \cdot (Y_1 - Y_0)$.

Definition 10. (Potential outcome) *The potential outcome for unit i given the treatment $t \in \{0,1\}$, written $Y_{t,i}$ is the value of Y_t for the unit i after it is treated as t. The observed outcome is $Y_i = Y_{0,i} + T_i \cdot (Y_{1,i} - Y_{0,i})$.*

The observed outcome is the one for the actual treatment of the unit. The counterfactual outcome is the one for the opposite treatment. In the kidney stones example, a patient (a unit) can either take the drug or not. Therefore, only one of the two potential outcomes is observed. The main problem in causal inference is to estimate the causal effect of the treatment over the outcome.

3 | CAUSAL DISCOVERY

In this section, we consider the problem of learning causal relationships among variables from observational data. We survey methods and tools to solve causal discovery and metrics to evaluate methods and public datasets to experiment with.

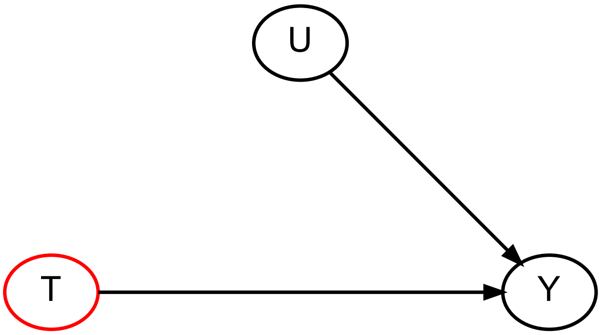


FIGURE 3 Example of a do-operation performed on the variable T in the kidney stones example

Depending on how a causal algorithm is constructed, it can be classified as either constraint-based or score-based. This type of classification is usually applied to Bayesian-like methods, but it can be extrapolated to other methods, provided they have a similar structure.

Constraint-based algorithms employ independence tests to identify a set of edge constraints for the graph using observational data, for example, using the G_2 test (Spirtes et al., 2000). Further rules determine then the direction of the found relationships. In exceptional cases, the rule phase is skipped to create undirected graphs. These graphs are usually local, meaning they only convey a particular node's (undirected) relationships.

Score-based algorithms assign a relevance score to candidate graphs through some adjustment measures, such as the Bayesian Information Criterion (BIC). However, these algorithms are computationally expensive since they have to enumerate (and score) every possible graph among the given variables. In addition, greedy heuristics are applied to restrict the number of candidates.

The section is structured by considering observational data, namely cross-sectional, time-series, and longitudinal data. The causal discovery methods, in fact, significantly depend on the type of data under analysis.

3.1 | Cross-sectional data

The search of causal relationships in cross-sectional data is one of the most investigated causal tasks.

> **Definition 11.** (Cross-sectional data) *Observation of subjects at one point or period of time, or for which the analysis has no regard to differences in time among the observations.*

This type of data is characterized by the fact that they are collected through the observation of several subjects simultaneously, this being not considered a study variable. These types of data are usually analyzed by comparing differences between subjects. Variables can have continuous, discrete, binary, or text data types. An excerpt is shown in Table 2.

The usage of this type of data for causal discovery has a significant downside. Since it represents a single point in time, it is not possible to exploit causal precedence (A causes B if A happens before B). This implies that to find the relationships' direction, it is necessary to apply an extra step. Various methods exist, covering all types of variables (binary, discrete, continuous, and mixed).

Perhaps the most known constraint-based causal discovery algorithm is PC (named after its authors, Peter and Clark; Spirtes et al., 2000). It relies upon the faithfulness assumption to create the models, meaning that all independencies must obey the d-separation criterion (Section 2). Like most constraint-based methods, this methodology consists of two phases: searching for (in)dependencies (also called *skeleton*[2] *phase*) and orienting dependencies.

In the first phase, the algorithm starts with a fully connected undirected graph. For each pair of adjacent variables A and B, it tests if the conditional independence $A \perp\!\!\!\perp B \mid C$ for a set C of variables is all adjacent to A (or, equivalently, all adjacent to B). Tests start with $C = \emptyset$ (unconditional independence) and iterate over sets of increasing size. If conditional independence holds, the undirected edge between A and B is removed.

TABLE 2 Excerpt from the *abalone* cross-sectional dataset

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16

The orientation phase applies a number of rules to direct edges (Spirtes et al., 2000):

1. Consider variables A,B,C such that $A-B-C$, namely A and B, B and C are adjacent, but A and C are not adjacent, that is, it holds in the skeleton phase that $A \perp\!\!\!\perp C \mid D$ for some D. If $B \notin D$, we orient the edges as $A \rightarrow B \leftarrow C$. The triple A,B,C is called a v-structure;
2. If there is a directed edge $A \rightarrow B$, and B and C are adjacent ($B-C$), but A and C are not adjacent, then $B-C$ is oriented as $B \rightarrow C$; and
3. If there is a direct path between A and B and an undirected edge between A and B, orient $A-B$ as $A \rightarrow B$.

PC-stable (Colombo & Maathuis, 2014) tackles a known problem inherent to PC known as order dependence. PC output depends on the order in which the variables are analyzed in the skeleton phase. This means that, if we have a $\text{order}_1(V) = \{A,B,C,D,E\}$ and $\text{order}_2(V) = \{A,D,B,E,C\}$, the resulting skeletons will not be the same. PC-stable tackles this by saving discarded nodes in a separate list instead of removing them right away, at each iteration (with a given size of C). The saved nodes are only removed permanently in the next iteration. This way, removing edges is no longer affected by the order of the independence tests at an iteration.

Another variant *conservative PC*. After creating the skeleton, this algorithm tests every potential v-structure $X-Y-Z$ by checking if $X \perp\!\!\!\perp Z \mid N$ where N includes all the neighbors of X and Z. If Y is not in all the separating sets or there are no variables in the set, $X-Y-Z$ is marked as *ambiguous*, and it is not directed. On the other hand, if Y is not in any separating set, the method continues as PC.

Although *PC* (and its variants) is a powerful tool to uncover causal relationships, it does not scale to high-dimensional data. For example, in the *PC-select* (sometimes called PC-simple) method (Bühlmann et al., 2010), the second phase is removed, and the conditional independence test is only applied to a target variable. Furthermore, because the method does not include an orientation phase, the output is an undirected graph.

Another strategy to tackle high dimensional data is to search for causal relations only locally to a target variable. The max–min parents and children algorithm (MMPC; Tsamardinos et al., 2006) adopts this approach using a Min-Max heuristic as a conditional independence test.

Although PC is considered as a benchmark algorithm for this type of data, it assumes causal sufficiency (Definition 7), meaning that it does not allow for open systems (systems with latent variables). For cases where the causal assumption cannot be fulfilled, *FCI* can be used (Spirtes et al., 1995). This method applies the same two phases of PC: the skeleton and orientation phases. In the skeleton phase, FCI applies a conditional independence test to find all the potential causal relationships. It is in the second phase that FCI differs the most from PC: instead of assuming that a relationship must have a direction (Glymour et al., 2019), the method tests possible d-separations $X \perp\!\!\!\perp Y \mid Z$ in the skeleton. If there is at least a variable in Z that d-separates the edge, then it is removed. After this, FCI applies several rules to direct the edges (Spirtes, 2001). FCI also differs from PC in the way it represents the relationships. Instead of two types of relationships (→ and −), FCI current implementations have four:

- $\mathbf{X} \rightarrow \mathbf{Y}$ that represents *X causes Y*;
- $\mathbf{X} \leftrightarrow \mathbf{Y}$ that represents that there is unmeasured confounders from both variables;
- $\mathbf{X} \circ\!\rightarrow \mathbf{Y}$ that represents either *X causes Y* or there is unmeasured confounders from both variables;
- $\mathbf{X} \circ\!-\!\circ \mathbf{Y}$ can represent: (1) *X causes Y*, (2) *Y causes X*, (3) there is unmeasured confounders from both variables, (4) *X causes Y* and there are unmeasured confounders from both variables or (5) *Y causes X* and there are unmeasured confounders from both variables.

The Anytime FCI is a slight modification of FCI that restricts the maximum number of variables in the separation set used to perform the conditional independence tests to a user-defined threshold.

The Adaptive Anytime FCI (Colombo et al., 2012) is similar to Anytime FCI in the way that it restrains the number of variables in the separation set. The critical difference is that, instead of the user defining this maximum, it is calculated by the algorithm, using $K = max_i(adj(C_1,X_i)|-1)$, in where C_1 represents the initial skeleton, X_i a vertice of C_1 and adj represents the list of adjacencies from X_i in C_1.

FCI and its variants can benefit from data preparation according to the Joint Causal Inference (JCI; Mooij et al., 2020) approach. This method extracts the context from several datasets, thus creating a pooled dataset where a traditional causal discovery method can be applied. This allows the generated model to encapsulate both information about the variables and the system from where these variables were measured. It is essential to understand that JCI is not a causal discovery method but a tool to prepare the data for it. The authors advocate its use with any causal discovery method but suggest its use with FCI specifically (hence FCI–JCI).

The Really Fast Causal Inference (RFCI; Colombo et al., 2012) is another FCI-like method that performs an additional test to the conditional independences before the v-structures phase: in this extra phase, the algorithm checks every unshielded triplet $X-Y-Z$ and examines $X \perp\!\!\!\perp Y \mid Z$ and $Y \perp\!\!\!\perp Z \mid X$. If this holds and Y is not in the separating set of X and Z, then this triplet is directed as $X \rightarrow Y \leftarrow Z$.

The RFCI-BSC (Jabbari et al., 2017) is a modification of RFCI, in where the Bayesian Scores Constraints (BSC) is used as a conditional independence test.

The Greedy Equivalence Search (GES; Chickering, 2002) is a score Bayesian-based method. It scales to high-dimensional data since it does not consider all existing patterns. This algorithm first adds new edges between two nodes X and Y, if these nodes are non-adjacent and there is no neighbor of Y that is not adjacent to X. Besides this, it also directs every edge of neighbor T of Y and not adjacent to X as $T \rightarrow Y$. Second, the method removes the best link in each iteration using the following criteria: it deletes every edge $X-Y$ or $X \rightarrow Y$ if there is a subset of neighbors of Y, Z that is adjacent to X. Besides, the algorithm transforms all edges $Z-Y$ as $Z \rightarrow Y$ and all edges $X-Z$ as $X \rightarrow Z$.

The Greedy Interventional Equivalence Search (GIES; Hauser & Bühlmann, 2012) is an improvement of GES. Besides adding and removing edges, this method has a third phase. In this phase, the algorithm elongates the DAG sequence, continuously modifying the original graph without altering the graph's skeleton. This new graph has the same number of edges and can be transformed into the original one by only changing one arrow.

The Fast Greedy Equivalence Search (FGS or FGES; Ramsey et al., 2017) is another modification of GES that uses parallelization to optimize the runtime of the algorithm.

The GFCI (Ogarrio et al., 2016) is a combination between the FGES and FCI. In this new method, both the skeleton and orientation phases of the pair are used: first, the skeleton phase of FGES is applied to the data, and then FCI is used to perfect the skeleton. The same happens in the orientation phase: initially, the algorithm accesses all the directed edges using FGES. This information is given to FCI, so it can use it to correct the edges' direction further.

3.1.1 | Software tools

The three most known tools/libraries for causal discovery in cross-sectional data are *pcalg*, *bnlearn*, and *Tetrad*.

Beginning with `pcalg` (Kalisch et al., 2012), this package has implementations of several causal methods, such as PC (original, conservative, and stable versions), GES, GIES, GDS, AGES, FCI (original, Anytime FCI, Adaptive Anytime FCI, and FCI–JCI, FCI+, and RFCI). Depending on the type of data used, this package offers default conditional independence tests for binary (G^2 test), discrete (G^2 test), and continuous (Fisher's z-transformation) data. Moreover, it is possible to adapt other conditional dependence tests to be used in this framework. For score-based methods (such as GES), *pcalg* includes the ℓ_0-penalized Gaussian maximum likelihood estimator for both discrete and continuous data.

`bnlearn` is a widely known and used R package (Scutari, 2010). This package provides an implementation for PC stable and MMPC, and it is possible to accommodate discrete, continuous, and mixed data by changing the conditional independence test. `Bnlearn` implements several conditional independence tests. For discrete data, `bnlearn` has the following tests available: mutual information (information-theoretic distance measure), shrinkage estimator for the mutual information (Hausser & Strimmer, 2009), and Pearson's χ^2 (classical version for contingency tables). For continuous data, the Pearson's linear correlation, Fisher's Z (transformation of the linear correlation with asymptotic normal distribution), mutual information (information-theoretic distance measure) and shrinkage estimator for the mutual information (Ledoit & Wolf, 2003) are available. Finally, for mixed data, mutual information (information-theoretic distance measure) is available.

Finally, `Tetrad` (Ramsey et al., 2018) is one of the most complete graphical tools for cross-sectional causal discovery. This tool implements the following methods: FCI, RFCI-BSC, FGES, GFCI, PC, and RFCI. `Tetrad's` methods can be applied in continuous, discrete, and mixed data by choosing the correspondent independence tests/score methods. For constraint-based algorithms, *Tetrad* also implements the following conditional independence tests. For discrete data, the conditional Gaussian test, χ^2 test, degenerate Gaussian likelihood ratio test, G^2 test, and probabilistic test are available. For continuous data, `Tetrad` presents the following tests: conditional correlation independence, conditional gaussian test, degenerate Gaussian likelihood ratio test, fisher Z test, and kernel conditional independence. Finally, the following tests are available for mixed data: conditional gaussian test and degenerate Gaussian likelihood ratio test. For score-based causal algorithms, *Tetrad* also offers several scoring methods. For discrete data, `Tetrad` offers the following tests: BDeu score, BIC score, conditional gaussian BIC score, and degenerate gaussian BIC score. For continuous data,

TABLE 3 Overview of software and methods for causal discovery in observational data

Software		Data				Type of algorithm			
		Categorical data	Continuous Data	Mixed Data	Time-series data	Causal Sufficiency	Constraint-based	Score-based	Non-Bayesian
bnlearn	MMPC	✓	✓	✓		✓	✓		
	PC	✓	✓	✓		✓	✓		
pcalg	AGES	✓	✓	✓		✓		✓	
	FCI	✓	✓	✓			✓		
	FCI-JCI	✓	✓	✓			✓		
	Anytime FCI	✓	✓	✓			✓		
	Adaptative Anytime FCI	✓	✓	✓			✓		
	FCI+	✓	✓	✓			✓		
	GDS	✓	✓	✓		✓		✓	
	GES	✓	✓	✓		✓		✓	
	GIES	✓	✓	✓		✓		✓	
	LINGAM	✓	✓	✓					✓
	PC	✓	✓	✓		✓	✓		
	CPC	✓	✓	✓		✓	✓		
	PC Select (PC simple)	✓	✓	✓		✓	✓		
	RFCI	✓	✓	✓			✓		
Tetrad	PC and PCStable	✓	✓	✓		✓	✓		
	CPC and CPCStable	✓	✓	✓		✓	✓		
	PcMax	✓	✓	✓		✓	✓		
	FGES/FGES-MB	✓	✓	✓		✓		✓	
	IMaGES	✓	✓	✓				✓	
	FCI	✓	✓	✓			✓		
	RFCI/RFCI-BSC	✓	✓	✓			✓		
	GFCI	✓	✓	✓				✓	
	MBFS	✓	✓	✓		✓	✓		
	GLASSO	✓	✓	✓					✓
	FOFC	✓	✓	✓		✓	✓		
	FTFC	✓	✓	✓					
	LiNGAM	✓	✓	✓					✓

Tetrad has CCI-score, extended BIC (EBIC) score, conditional gaussian BIC score and degenerate gaussian BIC score. Finally, conditional gaussian BIC score and degenerate gaussian BIC score are available for mixed data.

A summary overview of these frameworks can be found in Table 3.

3.1.2 | Datasets

We present here a few[3] key datasets used in causal discovery. These datasets originate in the context of classification or regression problems.

The *LUCAS* (LUng CAncer Simple set) benchmark synthetic dataset is a binary dataset, proposed as a toy example for the challenges created by the Causality Workbench project (Guyon et al., 2011) and is constituted by 12 binary variables, 2000 instances and represents 12 different causal relationships. This dataset represents several potential factors to the development of lung cancer and other unrelated factors. Because of its design and consequently causal properties, this dataset can be used to evaluate methods in terms of forecasting and terms of generated patterns. This dataset was adopted in Ramanan and Natarajan (2020) to study how context-specific independencies can be used to learn causal algorithms. A sample from this dataset is shown in Figure 4.

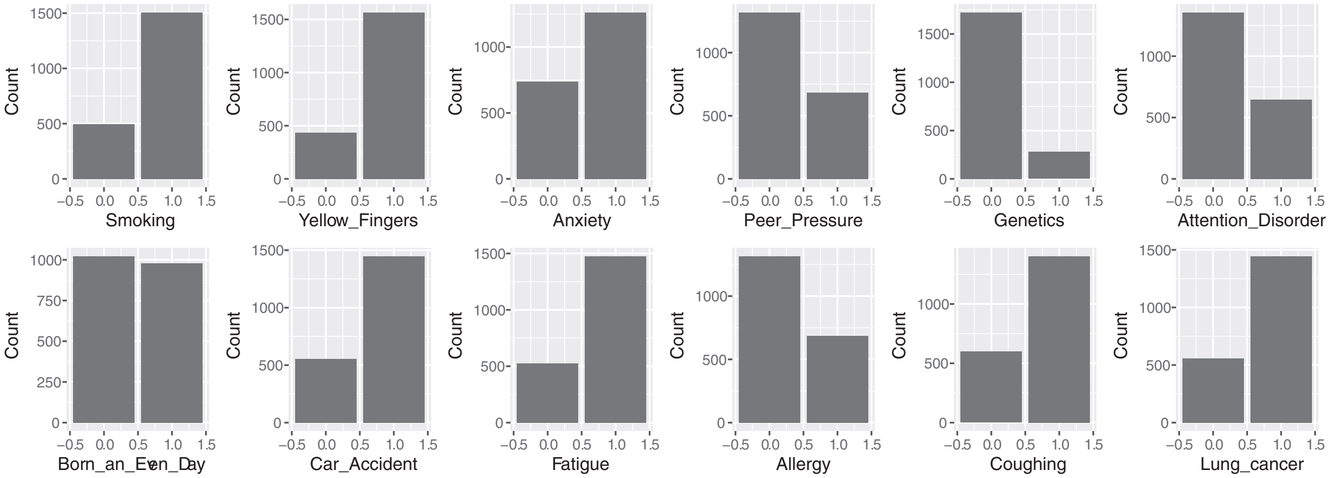


FIGURE 4 LUCAS' dataset distribuition

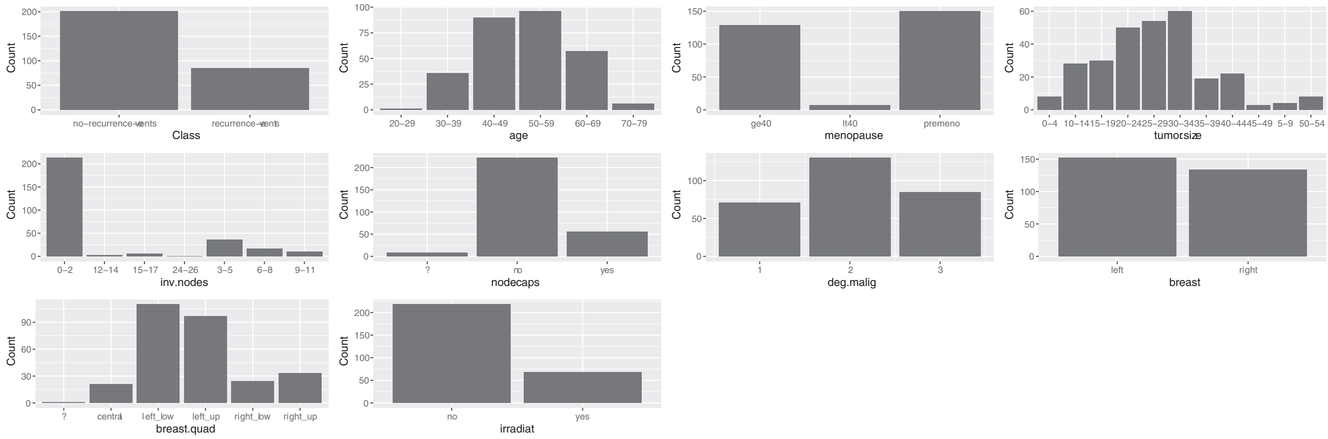


FIGURE 5 Breast cancer's dataset distribution

The *SIDO* (SImple Drug Operation mechanisms) is a real-world binary dataset representing a set of molecule descriptors tested against AIDS HIV and is constituted by 4932 variables and 12,678 instances. The authors did not make available any representation of the causal relationships present in the data. This dataset is part of a set of challenges proposed by the Causality Workbench project. The purpose of SIDO is to discover the causes for molecular activity in the descriptors. This dataset was used by Yu et al. (2021) to study causal feature selection methodologies.

The *Causal Protein-Signaling Networks in Human T Cells dataset*, sometimes called the Sachs dataset (Sachs, 2005), is a widely used dataset that represents proteins and phospholipids present in human immune system cells and is constituted by 11 discrete variables ($\{1,2,3\}$), 5400 instances, and represents 17 different causal relationships. This dataset aims to discover potential connections between the molecules without the need for physical intervention on them. For example, the comparative analysis of causal discovery algorithms in Singh et al. (2018) relies on this dataset.

The breast cancer dataset is a well-known discrete data set that entails information about cancer patients. This dataset is composed of 286 different instances and 9 variables, and its purpose is to diagnose the recurrence of breast cancer (Figure 5). Regarding causal discovery, this data was used by Dhir and Lee (2020), prove the efficiency of the algorithm proposed by them.

Finally, the asia dataset (Lauritzen & Spiegelhalter, 1988), is a widely used synthetic dataset that represents the relation between tuberculosis, lung cancer and bronchitis, and visitations to Asia (this dataset will be analyzed in more detail in Section 3.1.4). This dataset was also used in the work of Ramanan and Natarajan (2020) to study how context-specific independencies can be used to learn causal algorithms.

TABLE 4 Pattern metrics used for cross-sectional causal discovery methods

Metric	Description
Missing edges	Number of edges that are present in the original model but not in the generated one
Extra edges	Number of edges that are present in the generated model but not in the original one
Incorrect adjacencies (undirected edges)	Number of undirected edges that are present in the generated model but not in the original one
Correct directed edges	Number directed edges present in the generated model that were correctly directed
Incorrect directed edges	Number directed edges present in the generated model that were incorrectly directed
Structural hamming distance	Sum of missing edges, extra edges, and incorrectly directed edges
Structural intervention distance	For each pair X and Y checks whether the parents of X in the generated model are a valid adjustment set (Pearl, 2009) in the true model. If it is, it is counted as a correct procedure. If it is not, it is counted as a mistake.
Adjacency precision	$\text{Adj Precision} = \frac{\text{Correctly predicted adjacencies}^{a}}{\text{Predicted adjacencies}^{b}}$
Adjacency recall	$\text{Adj Recall} = \frac{\text{Correctly predicted adjacencies}}{\text{True adjacencies}^{c}}$
Arrowhead precision	$\text{Arrhd Precision} = \frac{\text{Correctly predicted arrowheads}^{d}}{\text{Predicted arrowheads}^{e}}$
Arrowhead recall	$\text{Arrhd Recall} = \frac{\text{Correctly predicted arrowheads}}{\text{True arrowheads}^{f}}$

[a] Number of undirected edges that are present in both the generated model and original one.

[b] All the edges found in the predicted model.

[c] All the edges found in the original model.

[d] Number of directed edges that are present in both the generated model and original one.

[e] All the directed edges found in the predicted model.

[f] All the directed edges found in the original model.

TABLE 5 Distribution for dataset *asia*

Number of attributes		8
Number of Instances		5000
Attribute	Yes	No
A	99.16%	0.84%
S	50.30%	49.70%
T	99.12%	49.70%
L	93.40%	0.88%
B	50.98%	49.02%
E	92.60%	7.40%
X	88.62%	11.38%
D	47.00%	53.00%

More examples can be found in the R packages bnlearn[4] and pcalg.[5,6]

3.1.3 | Evaluation metrics

Several metrics are used to evaluate causal discovery methodologies. These metrics are usually called *pattern metrics* as they search for common patterns between the ground-truth model that explains the data (or from which the data was generated) and the model generated by the method. Since generally, the ground truth model is represented in network form (DAGs, for example), these metrics are also related to network metrics. Despite this restriction, some models generated by non-Bayesian methods can be transformed into networks as long as the generated model is a rule-like model

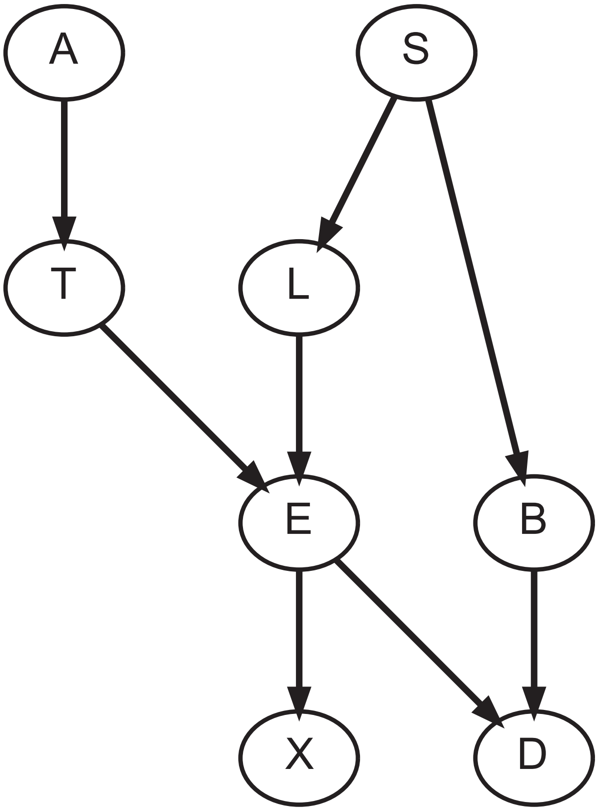
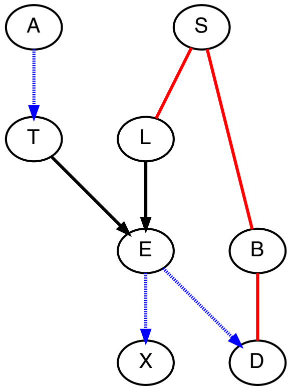


FIGURE 6 Ground-truth graph for dataset *asia*



(a) Network generated by pcalg's, bnlearn's andTetrad's PC

Metric	*PC*
Missing Edges	3
Extra Edges	0
Incorrect Adjencies	0
Correct Directed Edges	2
Incorrect Directed Edges	3
SHD	6
SID	30

(b) Pattern metrics for pcalg's, bnlearn's andTetrad's PC

Metric	*PC*
Adj Precision	100%
Adj Recall	62.50%
Arrhd Precision	40%
Arrhd Recall	25%

(c) Pattern metrics for pcalg's, bnlearn's andTetrad's PC

FIGURE 7 Comparison between the generated models and the true network for dataset *asia*. (a) Network generated by `pcalg's`, `bnlearn's`, and `Tetrad's` PC. Represents the missing edges (– or →), edges incorrectly directed or extra edges (– or →), and edges directed correctly (→). All three algorithms generated equivalent graphs and pattern metrics. (b) Represents the following metrics: Missing edges, extra edges, incorrect adjacencies, correct directed edges, SHD, and SID. (c) Represents the following metrics: Adjacency precision and recall, and arrowhead precision and recall

(such as association rule models) and given that all the generated relationships are simple (e.g., rules such as $\{A,B\} \rightarrow \{C\}$ are not allowed). Table 4 reports a collection of pattern metrics (Raghu et al., 2018).

TABLE 6 Excerpt from the *air quality* time-series dataset

Date	Time	CO (GT)	PT08. S1 (CO)	NMHC (GT)	C_6H_6 (GT)	PT08. S2 (NMHC)	NO_x (GT)	PT08. S3 (NO_x)	NO_2 (GT)	PT08. S4 (NO_2)	PT08. S5 (O_3)	T	RH	AH
March 10, 2004	18.00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
March 10, 2004	19.00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
March 10, 2004	20.00	2.2	1402	88	9	939	131	1140	114	1555	1074	11.9	54	0.7502
March 10, 2004	21.00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11	60	0.7867
March 10, 2004	22.00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
March 10, 2004	23.00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
March 11, 2004	00.00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603

TABLE 7 Excerpt from the *National Footprint Accounts 2018* longitudinal dataset

Country	ISO alpha-3 code	UN region	UN subregion	Year	Record	Crop land	Razing land	Forest land	Fishing ground	Built up land	Carbon	Total	Percapita GDP (2010 USD)	Population
Armenia	ARM	Asia	Western Asia	1992	BiocapPerCap	0.16	0.14	0.08	0.01	0.03	0	0.43	949.03	3,449,000
Armenia	ARM	Asia	Western Asia	1992	BiocapTotGHA	555,812.97	465,763.33	289,190.66	47,320.22	116,139.60	0	1,474,226.80	949.03	3,449,000
…	…	…	…	…	…	…	…	…	…	…	…	…		
Armenia	ARM	Asia	Western Asia	2014	EFProdPerCap	0.35	0.17	0.20	0.0006	0.062	0.62	1.40	3827.34	3,006,000
Armenia	ARM	Asia	Western Asia	2014	EFProdTotGHA	1,062,873.66	516,394.76	595,089.72	1692.15	185,046.34	1,856,992.85	4,218,089.49	3827.34	3,006,000
Afghanistan	AFG	Asia	Southern Asia	1961	BiocapPerCap	0.54	0.68	0.07	0	0.03	0	1.32		9,165,000
Afghanistan	AFG	Asia	Southern Asia	1961	BiocapTotGHA	4,990,784.71	6,212,850.07	654,431.08	0	272,261.57	0	12,130,327.43		9,165,000
…	…	…	…	…	…	…	…	…	…	…	…	…		
Afghanistan	AFG	Asia	Southern Asia	2014	EFProdPerCap	0.25	0.18	0.06	4.86e−05	0.05	0.11	0.65	610.24	31,628,000
Afghanistan	AFG	Asia	Southern Asia	2014	EFProdTotGHA	7,960,359.55	5,704,672.32	1,920,868.33	1536.006	1,458,818.88	3,372,775.04	20,419,030.13	610.24	31,628,000

Whenever a ground-truth model is not available, causal discovery methods can be evaluated regarding their performance in classification or regression tasks. In these cases, the traditional classification performance metrics are adopted (Hossin & Sulaiman, 2015).

3.1.4 | Running example

We elaborate an illustrative example to clarify the functionalities of libraries and the evaluation metrics. We restrict to the PC algorithm as implemented by `pcalg`, `bnlearn`, and `Tetrad`. The dataset under analysis is *asia*,[7] a discrete synthetic dataset, consisting of eight binary variables, each one with values *yes* and *no* (Table 5). This dataset was generated from the graph of Figure 6, which represents all the relationships present in the data.

Beginning with `pcalg`, since *asia* is composed of binary variables, we will use the stable version of PC (PC-stable) with the binary G^2 test. The implementation of PC in `bnlearn` can reason on discrete, continuous, and mixed data by changing the conditional independence test. For this example, we propose using mutual information as an independence test (others can be applied). Finally, the PC implementation available in `Tetrad` is similar to the previous two ones. For this example, we propose the usage of χ^2 test as an independence test.

The graphs generated by these three frameworks are equivalent and can be seen in Figure 7a.

To understand the evaluation metrics, we compare Figure 7a with the ground-truth graph (Figure 6), highlighting: *missing edges*, *extra edges*, *incorrect adjacencies*, *correct directed edges*, and *incorrect directed edges*. In all cases, there are three missing edges ($A \rightarrow T$, $E \rightarrow X$, and $E \rightarrow D$), and three incorrectly directed edges ($S \rightarrow L$, $S \rightarrow B$, and $B \rightarrow D$). The remaining edges in Figure 7a are correctly directed. Finally, there are no extra edges.

The structural hamming distance (SHD) is obtained by summing the missing edges, extra edges, and incorrectly directed edges. The structural intervention distance (SID) reflects how a mistake in the generated graph can influence the effects obtained. A summary of these metrics for the example at hand is shown in Table 7b. In addition, the derived performance metrics associated with the patterns in the graphs (adjacency and arrowhead precision and recall) are also shown. These metrics are a mixture of the more traditional prediction metrics, like precision or recall, but applied to evaluate patterns in graphs.

The adjacency and arrowhead precision and recall are calculated as follows. The adjacency precision is obtained by dividing the number of correctly predicted undirected edges (the number of edges with lines → and – or → in Figure 7a) by the number of predicted edges (Correctly predicted edges + extra edges): $\frac{5}{5} = 1 = 100\%$. The adjacency recall is obtained by dividing the number of correctly undirected edges by the number of true undirected edges (number of edges in the original graph): $\frac{5}{8} = 0.625 = 62.50\%$. The arrowhead or directed edges precision, is calculated by dividing the number of correctly predicted directed edges (the number of edges with → in Figure 7a) by the number of predicted directed edges (Correctly predicted directed edges + extra directed edges): $\frac{2}{5} = 0.4 = 40.00\%$. Finally, the arrowhead or directed edges recall, is calculated by dividing the number of correctly predicted directed edges by the number of true directed edges (number of directed edges in the original graph): $\frac{2}{8} = 0.25 = 25.00\%$.

It is worth pointing that the implementations of `pcalg`, `bnlearn`, and `Tetrad` are equivalent. Choosing one framework over the other depends on the user's requirements. For instance, if a user is not used to programming, `Tetrad` is a natural choice, given its user-friendly interface. If the user needs a wide range of causal methods and the possibility to use any conditional independence test (even the tests not available in the package), `pcalg's` is the best choice. However, if the user only wants to apply the library methods in an easy and fast way, `bnlearn` is the best choice.

3.2 | Time-series data

Time-series data include a sequence of observations about a single subject over multiple times.

Definition 12. (Time-series data) *Observations about a single subject at multiple points or periods of time, indexes in time order. We write X_t for the observation of random variable X at time t.*

This type of data is characterized by the fact that they are being collected in adjacent time periods, and there may be a correlation between distinct observations. Data collected on a continuous basis usually does not fall under the assumptions of conventional statistical methods, thus requiring different methods and tools. These types of data be univariate (only one variable is measured) or multivariate (multiple variables are measured) and the variables can be continuous, discrete, binary, text, among other types, as seen in Table 6.

The search of causal relationships among variables in time-series data has seen an exponential increase in interest in recent years, with sequential data collection becoming a common practice. Causal discovery from this type of data can overcome the problems found in cross-sectional data. Furthermore, since there is a time component, we can assume causal precedence: events in the present cannot cause events in the past. Thus, when faced with an identified (undirected) dependence, it is safe to assume the relationship's direction as past → future.

Several methods are specifically designed to solve the task of finding causal relationships in sequential observational data. One of the most known frameworks is the Granger causality, proposed by Granger (1969). Intuitively, X Granger-causes Y if predicting Y based on past observations and the past observations of X performs better than predicted Y based on its past only. Mathematically, this relationship can be formalized by testing that in the auto-regression:

$$Y_t = \sum_{j=1}^{m} a_j Y_{t-j} + \sum_{j=1}^{m} b_j X_{t-j} + \varepsilon_t$$

the coefficients b_j's are statistically significant.

In this equation, m represents the model order or the maximum number of lags to be used, a_j's and b_j's are the contributions of the delayed observation of Y and X, respectively.

More recent approaches include TsFCI (Entner & Hoyer, 2010), which is an adaptation of FCI for time-series data. This method uses sliding windows to transform the original time series into different subsets of consecutive timestamps, disregarding the time component in each subset and treating them as cross-sectional. The method creates a model for each subset of data using the models from previous timestamps as prior knowledge. Besides this, if a relationship disappears from the model m_t, this relation will be disregarded in the latter timestamps.

The PCMCI (Runge et al., 2019) is a causal graphical method designed to deal with linear and nonlinear time series. This algorithm is divided into two phases, each one corresponding to a different conditional independence test: the PC1 and MCI phases. In the PC1 phase, the algorithm applies the conditional independence strategy implemented by PC (skeleton phase) to uncover potential dependencies between each variable, in a specific timestamp, and all the other variables, in all the previous timestamps, for example, $X_t \perp\!\!\!\perp Y_{t-1} \mid Z$, $X_t \perp\!\!\!\perp Y_{t-2} \mid Z$, and so on, where t is the specific timestamp. Next, the method applies the MCI (momentary conditional independence) test (Runge et al., 2019) to further determine causal relationships between variables in different timestamps while taking into account autocorrelation and incorrect edge detections. PCMCI+ (Runge, 2020) is an extension of PCMCI, which admits the existence of contemporaneous links (a causal relationship between variables in the same timestamp). Because of this, PCMCI+ divides the skeleton search by type of relationships, namely lagged and contemporaneous relationships are found separately. LPCMCI (Gerhardus & Runge, 2020) is yet another PCMCI extension specifically designed to deal with latent variables. This method uses an FCI-like approach to represent the latent variables that are present in the relationships.

Time-series data is a particular case of longitudinal data (Definition 13; McArdle & Nesselroade, 2014, Chapter 1).

Definition 13. (Longitudinal data) *Observations about several subjects at multiple points or periods of time, indexes in time order, and subject.*

This type of data is characterized by the collection of information about the same individual at different points in time. This means that, for each subject in a dataset, there is a set of time-series variables that characterize him. The variables in longitudinal data can be continuous, discrete, binary, text, among other types, as seen in Table 7.

3.2.1 | Software tools

There are several libraries offered in different programming languages to solve the task of finding causal relationships in time-series data.

TABLE 8 Overview of software and methods for causal discovery in time series data

	Data				Type of algorithm				
Software		**Categorical data**	**Continuous data**	**Mixed data**	**Time-series data**	**Causal sufficiency**	**Constraint-based**	**Score-based**	**Non-Bayesian**
Tetrad	TsFCI	✓	✓	✓	✓		✓		
	TsGFCI	✓	✓	✓	✓			✓	
	TsIMaGES	✓	✓	✓	✓			✓	
	MultiFASK				✓			✓	
Tigramite (Runge et al., 2019)	PCMCI				✓		✓		
	PCMCI+				✓		✓		
	LPCMCI				✓		✓		
lmtest (Zeileis & Hothorn, 2002)			✓		✓				✓
NlinTS (Hmamouche, 2020)			✓		✓				✓

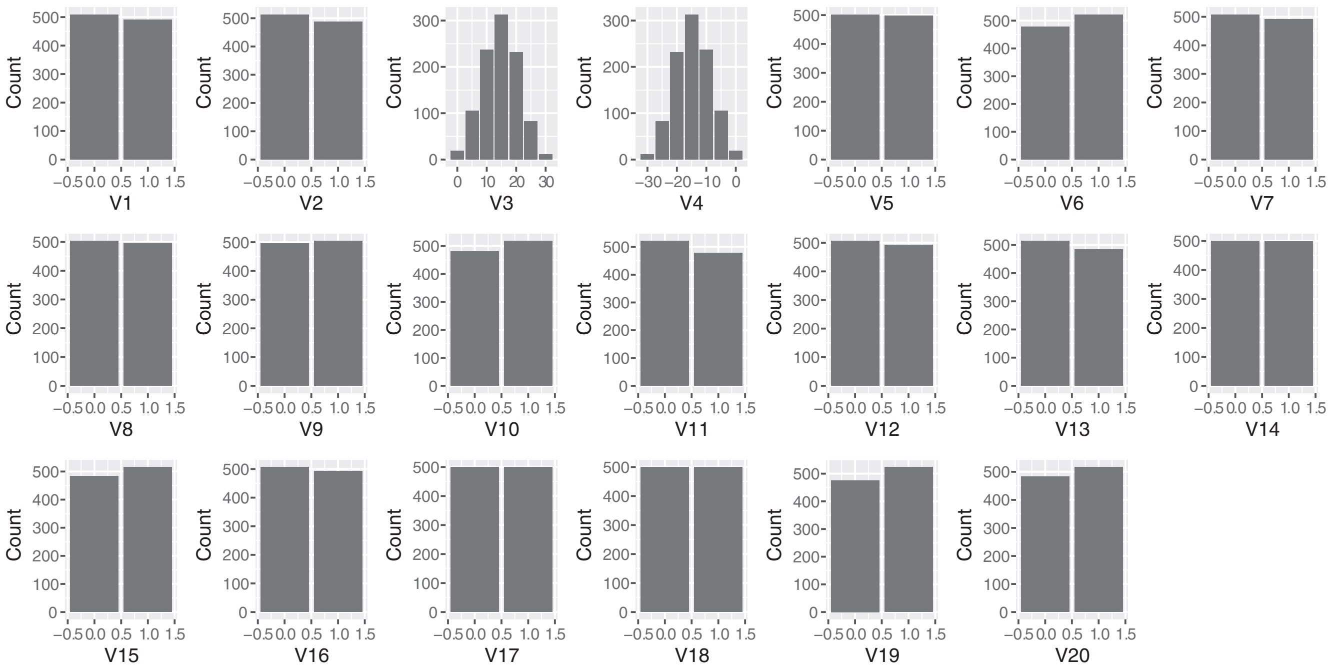


FIGURE 8 Distribution for one of the time-series in dataset FLAIRS

TABLE 9 Pattern metrics used in causal discovery from time-series data

Metric	Description
Accuracy	$\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{False Positive}+\text{False Negative}+\text{True Negative}}$
Mean/median error	Measures the differences between the predicted and ground truth. In this category, we can have all the variances of mean and median measures (root, squared, etc.)
Longest common subsequence	Measures the size of the longest sequence of events in a time-series model
Edit distance with real penalty	Measure the number of changes to transform one series into another, with a user-defined penalty
Euclidean distance	Measures the distance between each step of the series $d_E\left(\vec{x},\vec{y}\right)=\sqrt{\left(\vec{x}-\vec{y}\right)\left(\vec{x}-\vec{y}\right)'}$
Dynamic time warping	Measures the distance between two sequences. Being a sequence of a set of time points, the distance between each point is measured using the euclidean distance

lmtest (Zeileis & Hothorn, 2002) is an R package known mainly by its implementation of the Granger causality, as well as the standard dataset *ChickEgg* that is used as an example later in Section 3.2.2.

`NlinTS` (Hmamouche, 2020) is another R package. Similar to `lmtest`, this package implements a version of the Granger causality. Besides this, `NlinTS` implements a nonlinear version of this test.

`Tetrad`, the tool presented in Section 3.1.1, has also implementations for several methods that deal with time-series data, including TsFCI, FASK, and TsGFCI.

`Tigramite` (Runge, 2004–2021) is a Python framework for causal discovery in time-series data. This tool implements three different causal discovery methods (PCMCI, PCMCI+, and LPCMCI) and the following conditional independence test (all these tests can be used together with the causal discovery methods): ParCorr (Yagoubi et al., 2018), GPDC/GPDCtorch (Székely et al., 2007), CMIknn (Runge, 2018), and CMIsymb (Runge, 2018).

Unlike the previous data types, to the best of our knowledge, there is no tool available at the moment to deal specifically with longitudinal data. However, a few theoretical frameworks have been proposed for this type of data. One such framework is the Causal Inference over Mixtures (CIM; Strobl, 2019). This method infers the causal structure by creating a mixture of DAGs, using the Global Markov Condition (Definition 5). Explicitly designed for longitudinal medical

TABLE 10 Dataset *ChickEgg*

(a) Dataset excerpt		Chicken		Egg
1930		468,491		3581
1931		449,743		3532
1932		436,815		3327
…		…		…
1983		364,584		5656
(b) Dataset information				
Number of attributes			**2**	
Number of Instances			54	
Mean ± SD	Chicken		419,504 ± 46,406.94	
	Egg		4986.46 ± 884.97	

```
#Chicken cause Eggs?
> grangertest(egg ~ chicken, order = 3, data = ChickEgg)
Model 1: egg ~ Lags(egg, 1:3) + Lags(chicken, 1:3)
Model 2: egg ~ Lags(egg, 1:3)
  Res.Df Df      F Pr(>F)
1     44
2     47 -3 0.5916 0.6238

#Eggs cause Chicken?
> grangertest(chicken ~ egg, order = 3, data = ChickEgg)
Model 1: chicken ~ Lags(chicken, 1:3) + Lags(egg, 1:3)
Model 2: chicken ~ Lags(chicken, 1:3)
  Res.Df Df     F   Pr(>F)
1     44
2     47 -3 5.405 0.002966 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CODE 1 Egg cause chicken? or chicken cause eggs?

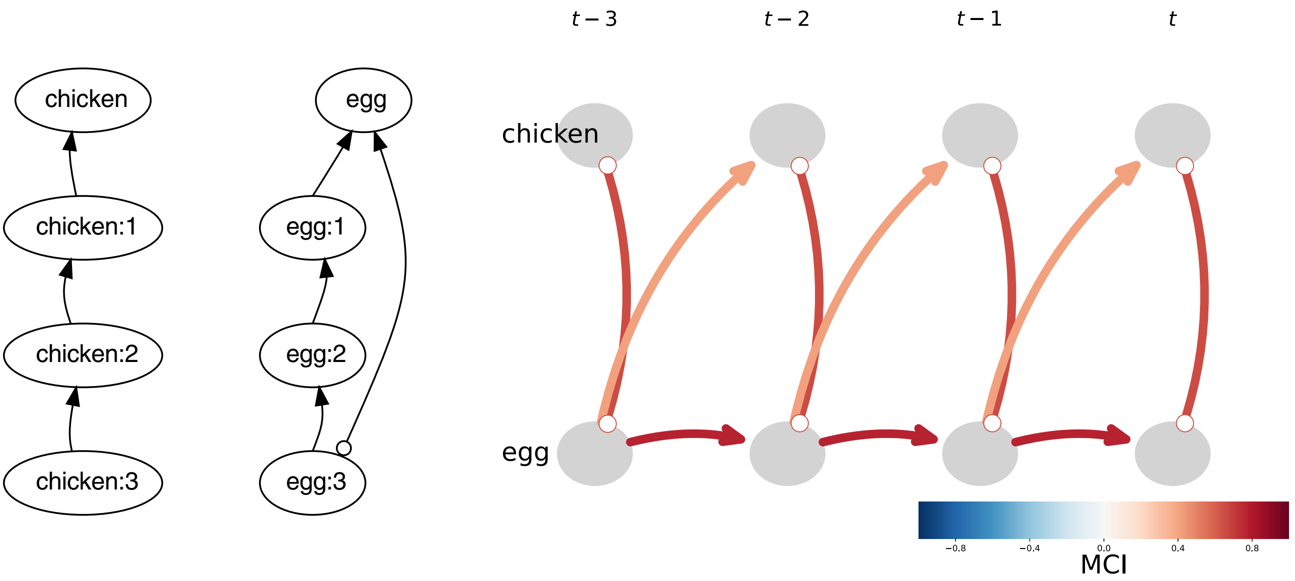
data, it allows for cycles. Besides this, it applies the skeleton phase of (Colombo & Maathuis, 2014). The orientation phase proposed by the authors is similar to FCI.

An overview of these libraries can be found in Table 8.

3.2.2 | Datasets

Compared to the case of cross-sectional data, there is a smaller range of benchmark datasets for time-series data[8] for times-series causal discovery algorithms is smaller.

The *FLAIRS 2015* (Huang & Kleinberg, 2015) dataset is synthetically generated. It comprises 22 different subsets, each with a different causal structure, with lags between 1 and 3, 20 continuous variables, and 1000 time-points each. The causal structures simulated in this dataset are common cause, common cause, and common effect, and random relationships. A sample from this dataset is shown in Figure 8 and Table 8.



(a) TsFCI

(a) PCMCI

FIGURE 9 Networks generated by *tsFCI* and *PCMCI* for dataset *ChickEgg*

TABLE 11 Performance metrics for dataset *ChickEgg*

	tsFCI	PCMCI
Mean squared error	19,009.5	17,491.32
Longest common subsequence	3	5
Edit distance with real penalty	1588.65	1563
Euclidean distance	515.88	494.85
Dynamic time warping	1001.87	736.42

The *FinanceCPT* (Kleinberg, 2012) is a simulated dataset composed of 25 portfolios (variables) with 10 causal structures each and 4000 day time periods. The structures simulated in this dataset are no dependency between portfolios, 20 random relationships with a lag one, 40 random relationships with a lag one, 20 random relationships with random lags 1–3, 40 random relationships with random lags 1–3, and many-to-one relationships at a lag of one.

The PROMO dataset is a time-series simulated dataset composed of a daily measurement of 1000 promotion variables and 100 product sales for 3 consecutive years. Its purpose is to identify which promotions affect sales. This dataset is part of a challenge proposed by the Causality Workbench project (Guyon et al., 2011).

The *ChickEgg* (Thurman & Fisher, 1988) is a time-series dataset with information collected annually about the number of chickens and eggs between 1930 and 1983. It consists of two variables (number of chickens and eggs per year) and three lags (this dataset will be analyzed in more detail in Section 3.2.4).

Regarding longitudinal datasets available for causal discovery, there are very few examples. Moreover, it is challenging to find ground-truth data to evaluate approaches because this area is relatively unexplored. One dataset is the *National Footprint Accounts 2018*, which collects data from the Ecological Footprint and biocapacity of countries across the world in over 50 years. The objective of this dataset is to understand the cause of the produced footprint values. A sample from this dataset is shown in Table 7.

3.2.3 | Evaluation metrics

The pattern metrics presented in Section 3.1.3 can be applied to time-series methods as well if there is a ground-truth model that represents the causal relationships present in the data. Table 9 shows a set of performance metrics specific to time-series data (Moraffah et al., 2021), to be used when this information is not available.

The *accuracy* is a metric used to evaluate classification models and can be defined as the fraction of correct predictions made by the model.

The *mean and median errors* are metrics that encapsulate the fraction of times the model got some response wrong. This error can be calculated in several ways, being the simplest one $1 - \text{accuracy}$.

The *euclidean distance* (Iglesias & Kastner, 2013) is another symmetric metric that calculates the distance between two-time series $\vec{x}$ and $\vec{y}$ (the predicted and the ground-truth). This metric is usually used for regression problems.

The *longest common subsequence* (Bagnall et al., 2016) is an asymmetric metric that measures the number of correct predictions in sequence and reports the highest number. This metric is usually used in regression problems since it uses the euclidean distance to calculate the difference between the predictions and the ground truth. This is performed by reducing the difference to 0 or 1 depending on the distance. If the Euclidean distance between two values is smaller than a defined threshold, they are considered equal. Hence, the distance is 0. On the other hand, if the difference is higher than the threshold, then the distance is 1.

The *Edit Distance with real penalty* (Chen & Ng, 2004) is another distance metric that reports the number of edits that are needed to transform the series of predictions into the ground truth.

Finally, the *Dynamic Time Warping* (Berndt & Clifford, 1994) is a distance metric that calculates the difference between two-time series, taking into account the potential differences in measurement in the timestamps (e.g., different frequencies). This is done by comparing each timestamp t from one of the time series with $t+1$, $t+2$, and so on, from the second time series.

With regard to metrics for evaluating causal methods for longitudinal data, there are two options. First, the evaluation metrics presented in Section 3.1.3 can be applied if there is a ground-truth structure to compare with. Second, since time-series data is a particular type of longitudinal data, the evaluation metrics presented in this section can also be applied.

3.2.4 | Running example

We elaborate an illustrative example to understand the libraries and evaluation metrics for time-series data. In this example, we consider the following methods: `lmtest's` Granger Causality, `Tetrad's` tsFCI, and `Tigramite's` PCMCI. The dataset under analysis is *ChickEgg*[9] (Table 10). There is no ground-truth graph, but it is well-known that from the dataset, it can be proven that eggs came before chickens (Egg → Chicken). To apply the evaluation metrics presented in Section 3.2.3, data is split into a training set (70%) and testing set (30%).

Beginning with `lmtest`, the Granger causality method needs no particular parameter insertion besides stating the number of lags. To be able to understand if *Eggs* cause *Chicken* or the other way around, it is necessary to apply the test in both cases: *Eggs cause Chicken* and *Chicken cause Eggs*). The output of the tests, reported in Code 1, shows that *Eggs cause Chicken* has a significant p-value (0.002996), while the other test has not. The Granger causality test is not a typical causal discovery method that creates a model that entails all the relationships in the data. Instead, to uncover relationships, it is necessary to test every combination of variables to find every relationship. Besides, this method reports only how the variables are related as a whole and not by lags (unlike the methods presented further in this section). Hence, it is impossible to compare it with the remaining ones further.

Another potential method to undercover causal relationships in time series data is `Tetrad's` tsFCI. In this method, it is possible to accommodate discrete, continuous, and mixed data. For the *ChickEgg* example, we use the Fisher's Z test as an independence test. The generated graph can be seen in Figure 9a.

Finally, PCMCI from `Tigramite` can also be used. This algorithm is also accommodates discrete and continuous data. For the *ChickEgg* example, we use the ParCorr test as an independence test. The generated graph is shown in Figure 9b.

Table 11 presents the evaluation metrics from Section 3.2.3. These metrics are calculated over the test set by comparing the original and predicted values of the target variable. The metrics *Edit Distance with Real Penalty*, *Euclidean Distance*, and *Dynamic Time Warping* are all distances, meaning that the lower the value, the better the performance is. The same applies to the *Mean Squared Error*. The *Longest Common SubSequence* represents the size of the largest sequence of values present in both the target's true values and predicted values. For this reason, the higher it is value is, the better performance the model achieves.

From Figure 9b, we can notice that PCMCI successfully finds relationships between the egg variable at time t and the chicken variable at time $t+1$. This is also reflected in the results from Table 11, where the PCMCI performance is better than those of the tsFCI.

It is worth noting that these metrics are not specific for causal discovery. Instead, they evaluate any classification or regression model. Since, first and foremost, a causal discovery model is a classification or regression model (depending on that used), it can be evaluated with such metrics.

4 | CAUSAL INFERENCE

Causal inference aims at estimating the causal effect of a specific variable (treatment) over a certain outcome of interest. Depending on the context of analysis, the causal effects can be quantified by different metrics, each focusing on different granularity levels. While we adhere here to the PO model, the same metrics can be re-stated in the context of the SCM framework (Aliprantis, 2015). In this section, we first present a number of causal effects of interest. Then, we discuss methods to estimate such causal effects starting from observational data, distinguishing whether or not they make the unconfoundedness assumption. Next, we cover software tools, datasets, and present a running example.

4.1 | Causal effects

At the highest granularity level, for each unit in the PO model, the Individual Treatment Effect models the effect of treatment on the unit.

Definition 14. (ITE) *The Individual Treatment Effect of unit i is*: $ITE_i = Y_{1,i} - Y_{0,i}$.

Despite recent efforts (Shalit et al., 2017; Yao et al., 2018), the estimation of ITE is challenging, as only the observed potential outcome is available in the data. At the population level, the Average Treatment Effect is the expectation of the ITEs. See Caron et al. (2020) for a review on ITE.

Definition 15. (ATE) *The Average Treatment Effect is*: $ATE = \mathbb{E}[Y_1 - Y_0]$.

The same definition can also be stated under the SCM framework. Assuming a binary treatment variable T, the Average Treatment Effect is equivalent to $\mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)]$.

In some contexts, the effect of interest can be restricted on the treated units, called Average Treatment Effect on Treated.

Definition 16. (ATT) *The Average Treatment Effect on Treated is*: $ATT = \mathbb{E}[Y_1 - Y_0|T=1]$.

The ATE might fail to capture causal effects due to the presence of heterogeneity among the units. This is overcome by the Conditional Average Treatment Effect.

Definition 17. (CATE) *Given a subset of covariates* $\boldsymbol{X}$, *the Conditional Average Treatment Effect for* $\boldsymbol{X} = \boldsymbol{x}$ *is*: $CATE(\boldsymbol{x}) = \mathbb{E}[Y_1 - Y_0|\boldsymbol{X} = x]$.

Notice that CATE boils down to ATE when $\boldsymbol{X}$ is the empty set.

The objective of causal inference is to estimate causal effects starting from observational data, where only factual outcomes are available.

Based on the type of available data and assumptions, we distinguish three main settings: experiments, observational data with unconfoundedness, and observational data with no unconfoundedness.

4.1.1 | Experimental data

In experimental settings, the treatment variable T is under the control of the researchers. Units are assigned either to a treatment group or to a control group. For instance, patients can either receive or not the drug.

The experiment is defined as a Randomized Controlled Trial (RCT) whenever the treatment assignment is performed randomly. This is the gold standard for estimating causal effects in social sciences and clinical trials. Randomization ensures that potential outcomes are independent of the treatment, that is, $\{Y_0, Y_1\} \perp\!\!\!\perp T$. This implies $\mathbb{E}[Y_t] = \mathbb{E}[Y_t | T = t]$ and then:

$$ATE = \mathbb{E}[Y_1 - Y_o] = \mathbb{E}[Y_1 | T = 1] - \mathbb{E}[Y_0 | T = 0] = \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0] \tag{2}$$

A further assumption is, however, needed to estimate causal effects. In the stable unit treatment value assumption (SUTVA), the observed outcome of a unit does not depend on the treatment assigned to other units (Cox, 1958), that is, $Y_i = Y_{T_i, i}$. From (2), the ATE can then be estimated starting from a sample of observed units as follows:

$$\widehat{\text{ATE}}_{\text{RCT}} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i$$

where n_t is the number of units i in the sample with $T_i = t$. An equivalent approach is given by running a simple OLS regression of Y_i given T_i. See Angrist and Pischke (2008, Chapter 2) for a comprehensive approach.

4.1.2 | Observational data with unconfoundedness

While RCTs are the gold standard for retrieving causal effects, they might be costly, unethical, or even impossible to run. In an observational setting, a number of assumptions is required to assess the causal effects directly from an i.i.d. sample of observations. The first one requires the independence of the potential outcomes from the treatment, not in general, but for a given a set of covariates.

Definition 18. (Unconfoundedness) *Unconfoundedness w.r.t. a set of covariates* $\boldsymbol{X}$ *holds if*: $\{Y_0, Y_1\} \perp\!\!\!\perp T \,|\, \boldsymbol{X}$.

Unconfoundedness is also referred to as ignorability, conditional independence (Lechner, 1999), or selection on observables (Barnow et al., 1980). Under the SCM framework, this can be assimilated to the backdoor criterion (Pearl, 2009). In the kidney stones example, if we know that people having larger kidney stones are more likely to receive the treatment compared to patients with smaller ones, then we can divide the population into smaller subgroups depending on the size of the kidney stones, and analyze each stratum as if the treatment is randomly assigned. The

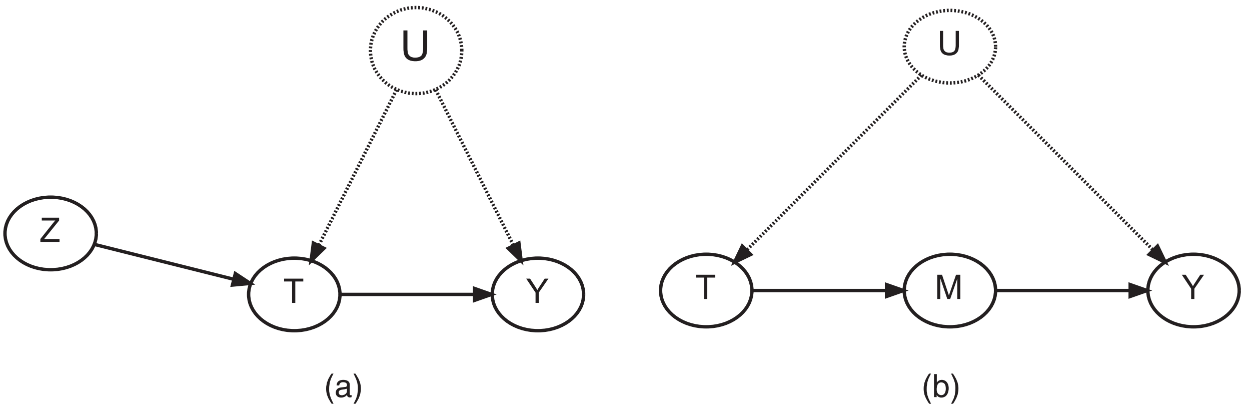


FIGURE 10 Valid examples for instrumental variables and Frontdoor criterion

additional assumption of the overlap condition (or positivity) requires non-deterministic treatment assignment, namely $0<P(T=1|\mathbf{X}=\mathbf{x})<1$. Unconfoundedness and overlap conditions together are known as strong ignorability. Assuming ignorability, an estimate of the ATE consists of averaging the estimates of CATEs over each stratum—a strategy known as stratification or blocking:

$$ATE=\sum_{\mathbf{x}}(\mathbb{E}[Y\mid T=1,\mathbf{X}=\mathbf{x}]-\mathbb{E}[Y\mid T=0,\mathbf{X}=\mathbf{x}])$$

In the literature, $\mathbb{E}[Y\mid T=1,\mathbf{X}=\mathbf{x}],\mathbb{E}[Y\mid T=0,\mathbf{X}=\mathbf{x}]$ are also referred to as conditional-response surfaces.

Such an approach has some issues in the case of continuous covariates or covariates with many values. An alternative relies on propensity scores (Rosenbaum & Rubin, 1983). Complete overviews of propensity scores methods for retrieving causal effects can be found in (Austin, 2011; Imbens, 2004; Imbens & Rubin, 2015).

Definition 19. (Propensity score) *For a set of covariates $\boldsymbol{X}$, the propensity score is the conditional probability of getting treated*: $e(\boldsymbol{x})=P(T=1|\boldsymbol{X}=\boldsymbol{x})$.

The estimation of propensity scores requires two key decisions: the model or functional form of $e(\cdot)$ and the covariates in $\mathbf{X}$. For binary treatments, the logistic regression model is commonly adopted (Caliendo & Kopeinig, 2008). Propensity scores are balancing scores, namely $\{Y_0,Y_1\}\perp\!\!\!\perp\mathbf{X}\mid e(\mathbf{X})$. Under the assumption of unconfoundedness, this implies that $\{Y_0,Y_1\}\perp\!\!\!\perp T\mid e(\mathbf{X})$ (Imbens & Rubin, 2015). Therefore, strata can be considered w.r.t. the propensity score, rather than on the covariates. This helps solve dimensionality concerns.

Another common usage of propensity score is matching (Abadie & Imbens, 2016) treated units with untreated units based on their similarity w.r.t. a distance measure either on the covariate space, or directly on the propensity score space (Yao et al., 2021). The idea is to pair individuals with different exposure to treatment, but similar in terms of the propensity score. Matching can be one to one, where each treated (resp., untreated) unit is paired with the closest untreated (resp., treated) unit, to one to many methods, where each unit can have multiple matches. By denoting with $J(i)$ the set of matched units for unit i, the ATE can be estimated from a sample of N units as:

$$\widehat{\mathrm{ATE}}_{\mathrm{MCT}}=\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{Y}_i(1)-\widehat{Y}_i(0)\right)$$

where $\widehat{Y}_i(t)=Y_i$ if $T_i=t$, and $\widehat{Y}_i(t)=\frac{1}{M}\sum_{j\in\mathcal{J}(i)}Y_j$ if $T_i=1-t$, where $M=|\mathcal{J}(i)|$.

Propensity score is also used to re-weight observational data (Hirano et al., 2003; Robins et al., 2000; Rosenbaum & Rubin, 1983) in such a way that the distribution of covariates is uniform across treatments, namely $w(\mathbf{x},1)\cdot P(\mathbf{X}=\mathbf{x}|T=1)=w(\mathbf{x},0)\cdot P(\mathbf{X}=\mathbf{x}|T=0)$. In the Inverse Probability of Treatment Weighting (IPTW; or Inverse Propensity Weighting [IPW]), the following weight definition achieves that:

$$w(\mathbf{x},t)=\frac{t}{P(T=t|\mathbf{X}=\mathbf{x})}+\frac{1-t}{1-P(T=1-t|\mathbf{X}=\mathbf{x})}=\frac{t}{e(\mathbf{x})}+\frac{1-t}{1-e(\mathbf{x})}.$$

Reweighting rebalances data to a sort of experimental data. As long as unconfoundedness and the overlap condition hold, the causal effects can be estimated as in the RCT scenario (but now for a weighted sample).

Some concerns regarding the variance of the estimators arise when units have propensity score values close either to 1 or 0. A popular solution is to stabilize the weights (Robins et al., 2000). Another methodology used for retrieving causal effects in this context is the G-formula (Robins, 1986). Both G-formula and IPW fall under the marginal structural models family (Petersen et al., 2006). For a comparison of marginal structural models techniques, see Chatton et al. (2020).

Since under unconfoundedness ATE can be characterized in terms of propensity scores and CATE, a natural question that arises is whether combining both approaches can have some benefits. This concern is addressed in (Robins et al., 1994), where Augmented Inverse Propensity Weighting (AIPW) is proposed. The AIPW estimator is defined as:

$$\widehat{\text{ATE}}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(\mathbf{x}_i)-\mu_0(\mathbf{x}_i)+T_i\cdot\frac{Y_i-\mu_1(\mathbf{x}_i)}{e(\mathbf{x}_i)}-(1-T_i)\cdot\frac{Y_i-\mu_0(\mathbf{x}_i)}{1-e(\mathbf{x}_i)}\right]$$

where the conditional-response surface $\mu_t(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t]$ and the propensity score $e(\mathbf{x})$ are estimated using regression models. The above estimator is also known as a doubly robust estimator, as it requires that only one of either the propensity score estimate or the response surface estimate is consistent to make the estimator consistent for the ATE (Scharfstein et al., 1999). Moreover, AIPW estimator is optimal among the non-parametric estimators, in the sense that it attains the Cramer–Rao bound (Hahn, 1998). See Glynn and Quinn (2010) for an overview of AIPW.

Another robust alternative is the Targeted Maximum Likelihood estimator method (Van Der Laan & Rubin, 2006), which involves four steps. In the first step, an estimation of the conditional expectation of outcomes given treatment and covariates $\mathbb{E}[Y \mid T, \mathbf{X}]$ is performed. In the second step, propensity scores are estimated. The third step adjusts the estimated conditional expectations using the estimated propensity scores. In the last step, the updated estimate of $\mathbb{E}[Y \mid T, \mathbf{X}]$ is used to generate pairs of potential outcomes and, for each unit, the difference between pairs of potential outcomes. The average of these differences is then an estimate of the ATE.

The methods presented assume homogeneous effects of treatment, which may not be the case in some contexts. Let us go back to the kidney stones example. In a (plausible) scenario where younger patients are more reactive to the treatment, the effectiveness of the drug will vastly change depending on the age of the patient. This is a case of heterogeneous effects. Here, the causal measure to focus on is the CATE. A semi-parametric approach for estimating CATE can be developed based on linearity assumptions [Chernozhukov et al., 2018; see Nie and Wager (2021) for a generalization to a nonlinear setting].

Another popular nonparametric approach is Bayesian Additive Regression Trees (BART; Hill, 2011), where the conditional expectation of the outcome given treatment and covariates is estimated by an ensemble (a sum) of decision trees, that is, $\mathbb{E}[Y_i|X_i = x, T_i = t] = f(x,t) = \sum_{q=1}^{Q} g_q(x,t)$ with $g_q(x,t)$ denoting a Bayesian decision tree. Therefore, the CATE $\tau(x)$ can be defined as $f(x,1) - f(x,0)$. In a similar fashion, other ensembles of trees and random forests have been investigated recently (Athey & Imbens, 2016; Athey & Wager, 2019; Wager & Athey, 2018). For a comprehensive review of these approaches, see Yao et al. (2021).

4.1.3 | Observational data without unconfoundedness

A practical problem for achieving unconfoundedness is to identify and collect all relevant covariates that make treatment assignments independent of the potential outcomes. A similar issue occurs for satisfying the overlap condition. Let us consider here approaches that do not assume unconfoundedness.

Instrumental variables (IV), first introduced by Wright (1928) and named by Reiersø (1945), were developed in the structural econometrics setting and were used to address endogeneity issues. For simplicity, assume no covariate $\mathbf{X}$, and consider a linear structural model $Y = \alpha + \beta \cdot T + \varepsilon$ of the outcome Y given the treatments T, with $\alpha = \mathbb{E}[Y_0]$ and $\varepsilon = Y_0 - \mathbb{E}[Y_0|T]$. Given this specification, unconfoundedness $(Y_t \perp\!\!\!\perp T)$ implies that $\mathbb{E}[\varepsilon|T] = \mathbb{E}[Y - \mathbb{E}[Y|T]|T] = \mathbb{E}[Y|T] - \mathbb{E}[Y|T] = 0$. In such a setting an OLS regression can consistently estimate the parameter β. Conversely, if unconfoundedness does not hold, it might be $\mathbb{E}[\varepsilon|T] \neq 0$, hence the estimation of the causal effect β through OLS regression is not consistent.

Other estimators to retrieve β when multiple instruments are available are the Two Stage Least Squares (2SLS; Theil, 1961) and the Method of Moments (Baum et al., 2003). In summary, the following conditions are required for using the IV approach: exogeneity condition: $Z \perp\!\!\!\perp \varepsilon$; exclusion restriction: the instrumental variable Z affects the outcome Y only through T; relevance condition: $Cov(Z,T) \neq 0$. The exclusion restriction can be easily understood through a graphical example, where no arrows pointing to Y are coming directly from Z, as shown in Figure 10a.

Modeling the treatment effect as a linear model can be considered a strong assumption. However, as shown by 2021 Nobel's price laureates in (Angrist et al., 1996; Imbens & Angrist, 1994), the IV strategy can also be used in the potential outcome framework where units might not comply with the designation of the treatment. Let us think, for example, of an experiment where the patients are randomly assigned to take a certain drug, but not all of them actually follow the assignment. In such a scenario, we can denote the outcome as Y_i, the treatment received as T_i (non-random) and the treatment to which each unit was assigned as Z_i (random). In this setting, we can frame the actual treatment T_i as a potential outcome of the assignment Z_i, that is, there are $T_{0,i}$ and $T_{1,i}$. Moreover, the outcome is now a potential

TABLE 12 Overview of software and methods for causal estimation

Software	Language		Estimands	Unconfoundedness					No Unconfoundedness				
	Python	R		Stratification	Matching	Weighting	DR	ML based	IV	Front door	RDD	DD	SC
dowhy (Sharma et al., 2019)	✓		ATE, ATT, ITE, CATE, LATE	✓	✓	✓	✓	✓	✓	✓	✓		
econML (Research, 2019)	✓		ATE, ATT, ITE, CATE, LATE					✓	✓				
Matching (Sekhon, 2011)		✓	ATE, ATT, CATE		✓								
MatchIT (Ho et al., 2011)		✓	ATE, ATT, CATE		✓								
R-FLAME (Orlandi et al., 2020)		✓	ATE, CATE			✓							
dame-flame (Dieng et al., 2019)	✓		ATE, CATE			✓							
PSW (Mao & Li, 2018)		✓	ATE, ATT, CATE			✓							
CBPS (Fong et al., 2021)		✓	ATE, ATT, CATE, LATE			✓	✓		✓				
ipw (van der Wal & Geskus, 2011)		✓	ATE, ATT, CATE			✓	✓						
PSweight (Zhou et al., 2021)		✓	ATE, ATT, CATE			✓							
RISCA (Foucher et al., 2020)		✓	ATE, ATT			✓		✓					
CausalGAM (Glynn & Quinn, 2017)		✓	ATE, ATT			✓	✓						
tmle (Gruber & van der Laan, 2012)		✓	ATE, ATT, CATE				✓						
BART (Sparapani et al., 2021)		✓	ATE, ATT, CATE					✓					
grf (Tibshirani et al., 2020)		✓	ATE, ATT, CATE					✓					
causalML (Chen et al., 2020)	✓		ATE, ATT, ITE, CATE, LATE					✓	✓				
CEVAE (Shalit et al., 2017)	✓		ATE, ITE, CATE					✓					
SITE (Yao et al., 2018)	✓		ATE, ATT, CATE					✓					
ivreg (Fox et al., 2021)			LATE						✓				
rddrobust (Calonico et al., 2021)			ATE at the cutoff								✓		
rddtools (Stigler & Quast, 2016)		✓	ATE at the cutoff								✓		
rdd (Dimmery, 2016)	✓		ATE at the cutoff								✓		
plm (Croissant & Millo, 2008)		✓	ATE									✓	
linearmodels (Sheppard et al., 2021)	✓		ATE									✓	
Synth (Abadie et al., 2011)		✓	ATE										✓
causalImpact (Brodersen et al., 2014)	✓	✓	ATE										✓

outcome of both T_i and Z_i and it can be defined (with a slight abuse of notation) as $\{Y_{t,z}\}_{t,z \in \{0,1\}}$. At this point, the three conditions of instrumental variables are met: exogeneity is ensured by the randomness of Z, the exclusion restriction holds as Z affects the outcome only through the actual assignment and, finally, the treatment assignment influences the actual taking of the drug, fulfilling the relevance condition. By looking at the actual treatment and the treatment assignment, we can identify four sub-populations in our sample: the compliers, who are those who follow the treatment designation, the always-takers, who are those that always receive the treatment no matter what is the assignment, the never-takers, who are those that never receive the treatment no matter what is the assignment, and the defiers, who are those that do the opposite of what they were assigned to. By adding the additional assumption of no defiers, the IV estimator $\hat{LATE} = \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[T_i|Z_i=1]-\mathbb{E}[T_i|Z_i=0]}$ identifies the treatment effect for those who comply with the assignment. Such an effect is also known as Local Average Treatment Effect (LATE). See Imbens (2014) for an overview of other instrumental variables approaches.

Another popular technique dealing with the lack of unconfoundedness is the front door criterion (Pearl, 1995) under the SCM framework. As depicted in Figure 10b, a few requirements need to be met in order to exploit it in the estimation of causal effects. In particular, M must block all the directed paths from T to Y, T has no unblocked backdoor paths to M and T blocks all the paths from Y to M. As a fact, M can be defined as a mediator variable. By estimating the effect of T on M and then of M on Y, it is possible to retrieve the effect of T on Y.

Another option arises whenever the treatment assignment depends only on the values of a specific variable. For instance, consider in the kidney stones example, that patients are assigned to treatment only if the size of their kidney stones is bigger than a certain threshold. The Regression Discontinuity Design (RDD; Thistlethwaite & Campbell, 1960) precisely assumes that the treatment depends on a variable C, called the running variable, and a threshold c_0 as, for unit i: $T_i = 1_{C_i > c_0}$. The core idea of this methodology is that observations close to the threshold are in principle pretty similar to each other, and they can be used to retrieve the causal effect of the treatment. In particular, two assumptions are made: (i) the probability of receiving the treatment jumps at the cutoff c_0:

$$\lim_{c \to c_0^-} P(T=1|C=c) = x^- \neq \lim_{c \to c_0^+} P(T=1|C=c) = x^+;$$

and, (ii) the potential outcomes are continuous at the cutoff, namely there exist:

$$\lim_{C \to c_0} \mathbb{E}[Y_0|C=c], \quad \lim_{C \to c_0} \mathbb{E}[Y_1|C=c].$$

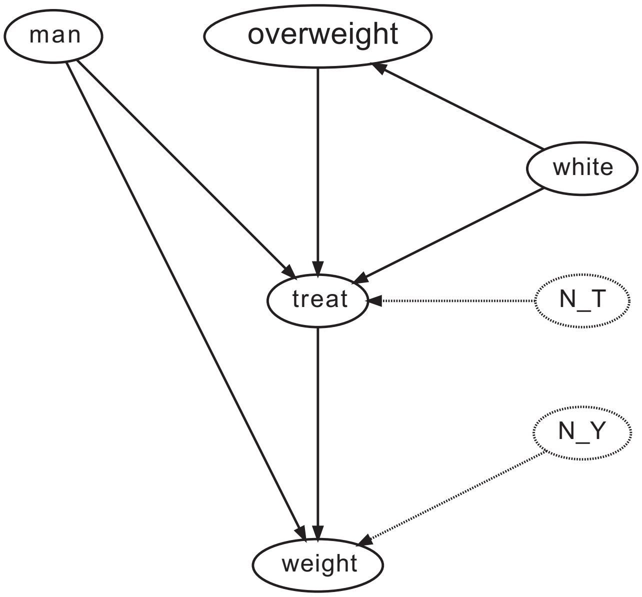


FIGURE 11 DAG for the running example on causal inference

Under such assumptions, the ATE can be calculated as:

$$\mathrm{ATE}_{\mathrm{RD}} = \mathbb{E}[Y_1 - Y_0 | C = c] = \lim_{C \to c_0^+} \mathbb{E}[Y | C = c] - \lim_{C \to c_0^-} \mathbb{E}[Y | C = c].$$

$\mathrm{ATE}_{\mathrm{RD}}$ can be estimated by either parametric or nonparametric approaches (Cattaneo et al., 2020, 2021; Lee & Lemieux, 2010).

In the case of longitudinal data, namely for each unit there are multiple observations over time, some further approaches are available. Let us consider a setting where the observed outcome $Y_{i,l}$ and a binary treatment $T_{i,l}$ are measured over time $l = 1, \ldots, L$ for each unit $i = 1, \ldots, N$ in a sample. Let us also suppose that the two variables are linearly related as: $Y_{i,l} = Y_{i,l}(0) + T_{i,l} \cdot \tau$, where $Y_{i,l}(0)$ refers to the potential outcome under no treatment at time l and τ can be seen as the constant causal effect of treatment T_i. Assuming that the effect is not heterogeneous, that is, the treatment affects the units in the same manner in all the periods, and that there are no treatment dynamics, that is, the outcome at time l depends only on the treatment at the same time, a causal estimate can be modeled as:

$$Y_{i,l} = \alpha_i + \beta_l + T_{i,l} \cdot \tau + \varepsilon_{i,l} \quad \mathbb{E}[\varepsilon | \alpha, \beta, T] = 0.$$

This two-way model assumes that there are two fixed effects, one for each unit and one for each period. In the simplest possible scenario, that is, when there are just two periods ($L = 2$), some units never get treated, and the remaining are treated just in the second period, the OLS estimator coincides with

$$\widehat{\mathrm{ATE}}_{\mathrm{DID}} = \frac{1}{|\{i : T_{i,2} = 1\}|} \sum_{i : T_{i,2} = 1} (Y_{i,2} - Y_{i,1}) - \frac{1}{|\{i : T_{i,2} = 0\}|} \sum_{i : T_{i,2} = 0} (Y_{i,2} - Y_{i,1})$$

$\widehat{\mathrm{ATE}}_{\mathrm{DID}}$ is what is referred to as a difference-in-differences (DID) estimator of the ATE. A seminal example of this technique can be found in Card and Krueger (1994), where the authors studied the effect of a rise in the minimum wage on unemployment. This was done by comparing New Jersey workers, where a minimum wage raise was introduced, with Pennsylvania workers, where no raise occurred. A crucial assumption of the difference-in-differences approach is the parallel trend, that is, although treatment and comparison groups may have different levels of the outcome prior to the start of treatment, their trends in the outcomes before the treatment should be the same. Moreover, as pointed out in Bertrand et al. (2004), extra care should be devoted when handling the error terms $\varepsilon_{i,l}$. For a comprehensive review of DID estimators, see Lechner (2011). In order to overcome the assumption of a fixed α_i over time, some approaches have been developed, such as the synthetic control methods, first introduced in Abadie and Gardeazabal (2003) and later developed in Abadie et al. (2010). The core idea of synthetic methods is to re-weight the units that were not treated so that the parallel trend assumption becomes more plausible. A more recent advancement in this direction is the synthetic difference-in-differences method, presented in Arkhangelsky et al. (2019), which also provides a unified perspective of DID and synthetic control methods.

TABLE 13 Mean and standard deviation of estimated ATEs over 1000 simulation runs

		Estimates				
	Ground truth coefficient	OLS regression	Exact matching	Full matching	Stabilized IPW	Doubly robust AIPW
High impact scenario	−30	−29.99 ± 0.766	−29.99 ± 0.822	−30.05 ± 6.92	−30.15 ± 0.839	−30.16 ± 0.841
Low impact scenario	−10	−9.993 ± 0.75	−9.989 ± 0.791	−9.79 ± 7.05	−10.14 ± 0.809	−10.159 ± 0.812
No impact scenario	0	−0.014 ± 0.736	−0.032 ± 0.776	0.265 ± 6.801	−0.183 ± 0.789	−0.201 ± 0.792

4.2 | Software tools

Many software tools for estimating causal effects have been developed in the recent years. We review a non-exhaustive list (Table 12), discussing which approaches they implement and their most relevant features.

DoWhy (Sharma et al., 2019) is one of the most complete tools. Written in Python, it provides a unifying framework for several methodologies, covering virtually the whole process of causal inference. `DoWhy` covers four tasks: model the causal problem through a causal graph, identify the causal estimand of interest, estimate the causal effect and validate the obtained results. The following identification strategies are currently implemented: backdoor criterion, front-door criterion, instrumental variables, and mediation analysis. For each of these, a few methods for estimating causal effects are provided. Moreover, `DoWhy` natively connects to the wide range of machine-learning-based estimators from `EconML` (Microsoft Research, 2019).

Propensity score matching is implemented in the following three tools. The popular R package `Matching` (Sekhon, 2011) offers three main functions: `Match`, `MatchBalance`, and `GenMatch`. The first one performs propensity score matching. The second one checks whether the matching method balances the data. The last function generates optimal weights for each covariate through a genetic search algorithm to automatically balance the data. `MatchIt` (Ho et al., 2011) is another popular choice. The package offers the function matchit, which generates matched data through different methods. After the matching is performed, standard regression techniques can be applied to the obtained data to retrieve the causal effects of interest. Another interesting option for large datasets with discrete covariates is the R package `R-FLAME` (Orlandi et al., 2020) and its Python version dame-flame. The packages provide a framework that implements the Fast, Large-Scale Almost Matching Exactly (FLAME; Wang et al., 2021) and Dynamic Almost Matching Exactly (DAME; Dieng et al., 2019) approaches.

Propensity score weighting is implemented in the following other four tools. `PSW` (Mao & Li, 2018), one of the main reference libraries, provides several techniques based on propensity score through the function `psw`. This package allows to check visually the propensity score distribution in both treatment groups, evaluate the covariates balance, and test the specification of the propensity score model. CBPS (Fong et al., 2021), available in R, implements several methods presented in Imai and Ratkovic (2013), both for cross-sectional data and longitudinal ones. CBPS maximizes at the same time covariate balance and the prediction of treatment assignment, while typically propensity score algorithms predict the treatment assignment and then perform a check on the covariates to see whether they are balanced among different treatment groups. This makes the method more robust to misspecifications. CBPS includes matching, weighting, and double-robust methods based on the estimated propensity scores. The package ipw (van der Wal & Geskus, 2011) implements the inverse probability of treatment weighting, both for time-fixed and time-varying frameworks. Similarly, `PSweight` (Zhou et al., 2021) covers propensity-score based estimators through Propensity Score weights, exact-matching weights, entropy weights, ATT weights, and overlap weights.

`RISCA` (Foucher et al., 2020) is a viable option for users interested in marginal structural models, as it provides functions for G-estimation and Inverse Probability Weighting.

Doubly robust estimators are offered also by `CausalGAM` (Glynn & Quinn, 2017), an R library that implements both standard estimators and the AIPW estimator, or by the `tmle` (Gruber & van der Laan, 2012) package, which provides an R implementation of the Targeted Maximum Likelihood Estimators.

Several tools rely on machine learning for estimating causal effects. `BART` (Sparapani et al., 2021) provides a package for the estimation of Bayesian Additive Regression Trees. `grf` (Tibshirani et al., 2020) is an R package that implements tree-based methodologies (Athey et al., 2019) for CATE estimation. The main function is `causal_forest`, which trains a causal forest to retrieve heterogeneous treatment effects. For Python, we refer to `EconML` and `Causal ML` packages. The former provides several implementations of state-of-the-art methodologies for retrieving heterogeneous causal effects, such as Double Machine Learning (Chernozhukov et al., 2018), causal trees/forests (Athey & Imbens, 2016; Wager & Athey, 2018), Doubly Robust Learning (Foster & Syrgkanis, 2020), and Meta-Learners (Künzel et al., 2019). The latter (Chen et al., 2020) covers similar approaches, yet less extensively. Further Python packages are available to estimate ITE, such as CEVAE (Shalit et al., 2017), and SITE (Yao et al., 2018), which follow the work of (Yao et al., 2018).

Instrumental variables estimation is offered by the R package `ivreg` (Fox et al., 2021). The function ivreg fits different estimators, including the Two-Stage Least Squares (2SLS) and the Method of Moments (MM).

`rdrobust` (Calonico et al., 2021) is one of the main tools offering an RDD implementation. It covers all the required steps through different functions: `rdplot` deals with the graphical exploration of the setting, `rdbwselect` picks the optimal bandwidth, and rdrobust computes the RDD estimator under different assumptions. Other packages for RDD estimation include (Stigler & Quast, 2016; Dimmery, 2016).

A canonical implementation of difference in differences can be obtained by simply instantiating a common panel linear regression [`plm` (Croissant & Millo, 2008) for R users, `linearmodels` (Sheppard et al., 2021) for Python researchers]. More advanced techniques, such as synthetic control, can be implemented through the `Sytnth` (Abadie et al., 2011) package and `CausalImpact` (Brodersen et al., 2014). The latter is available both in R and in Python.

4.3 | Datasets

We present a few open-access datasets widely used in papers about causal inference.

IHDP is based on the RCT developed by the Infant Health and Development Program (Brooks-Gunn et al., 1992). There are 25 covariates about infants and their mothers. The treatment regards the access to comprehensive early intervention. The goal of the study was to understand how this extra care could help reduce the developmental and health problems of low birth weight for premature infants.

Lalonde is a well-known observational dataset used to evaluate propensity score matching in (Dehejia & Wahba, 1999). The variables refer to workers' characteristics, such as age, education, ethnicity, marital status, exposure to the training program (the treatment), and salary.

Several datasets can be obtained from the Wooldridge package in R. This library includes all the 114 datasets of Wooldridge (2015). For instance, the *mathpnl* dataset (Papke, 2005) regards student performances at schools.

Some data can be obtained exploiting the popular RDD packages previously described. In particular, the *house* dataset from Lee (2008) and also used in Imbens and Kalyanaraman (2012), is included in the rddtools library. The dataset refers to observations for elections, and it was used to estimate the effect of being the incumbent, exploiting the percentage of votes as a running variable.

The *Card–Krueger 1994* dataset is a notorious example for DID estimation from (Card & Krueger, 1994; see also the version from Ropponen (2011)). It contains data about workers in the fast-food industry.

Finally, additional datasets from the econometrics domain are included in the R-package `AER` (Kleiber & Zeileis, 2008). These datasets can be used to estimate several causal inference techniques. For instance, the `Cigarettes` dataset is suited for instrumental variable approaches.

4.4 | Running example

A toy example is provided here to compare the performance of several methodologies under the unconfoundedness assumption. In such an example, we are going to generate data that simulate patients that undergo a diet. The code used for the whole experiment can be found through the link here 1. The structure of the data is illustrated in Figure 11.

For each instance, seven variables are simulated. The binary variable *man* is equal to one if the patient is a man, and it is generated through a $\mathrm{Bernoulli}(0.5)$. The binary variable *white* is equal to one if the patient is white, and it is distributed as a $\mathrm{Bernoulli}(0.65)$. The binary variable *overweight* is equal to one if the unit was overweight in the past, and it follows a $\mathrm{Bernoulli}(0.32)$ if *white* is equal to one, otherwise, it follows a $\mathrm{Bernoulli}(0.4)$. The variables N_T and N_Y are noise variables distributed as $\mathrm{Normal}(4,2)$ and $\mathrm{Normal}(15,1)$ respectively. The binary variable *treat* refers to whether the unit has been on a diet and it is defined as follows: if $(N_T + \mathrm{white} + 3 * \mathrm{overweight} - \mathrm{man}) > 6$, it is equal to one, otherwise it is equal to zero. Finally, the outcome variable weight, measuring the patient's weight, is simulated as follows: if the unit is a man, then $\mathrm{weight} = \mathrm{Normal}(200,30) + \mathrm{treat} \cdot \tau + N_Y$, otherwise $\mathrm{weight} = \mathrm{Normal}(150,30) + \mathrm{treat} \cdot \tau + N_Y$. In such a setting, τ captures the causal effect of the impact of *treat* over *weight*. Let us notice that the relationship between treat and weight is assumed to be linear and that unconfoundedness holds given *man*, *white*, and *overweight*.

Therefore, the goal of the whole example is to estimate τ. Three different scenarios are simulated: one where the impact is high ($\tau = -30$), one where it is smaller ($\tau = -10$), and one where there is no effect ($\tau = 0$).

For each scenario, 1000 simulations were run. In each simulation, a dataset containing 10,000 observations is generated and an estimate of the ATE is computed using the following methods: OLS linear regression, exact (one-to-one) matching estimator, full matching, stabilized IPW, and doubly robust AIPW.

The OLS regression coefficient is retrieved by running a linear regression of *weight* over *treat*, *man*, *white*, *overweight*, and the interactions of the centered-around-the-mean versions of *man*, *white*, and *overweight* with *treat*. For exact and full matching, `MatchIT` package is used. In exact matching, a complete cross of the covariates is used to form subclasses defined by each combination of the covariate levels. If a subclass does not contain both treated and control units is discarded. In this way, the remaining subclasses contain treatment and control units that are exactly equal in the included covariates. On the other hand, in full matching, both treatment and control units (i.e., the "full" sample) are assigned to a subclass and receive at least one match. In this case, the sum of the absolute distances between the treated and control units in each subclass is as small as possible. Stabilized IPW estimates are found using the `ipw` package, while the `CausalGAM` implementation is used to retrieve the Doubly Robust AIPW estimate.

Table 13 shows the mean of the estimated ATEs for each methodology over the 1000 runs and the relative standard deviations. Comparing the estimated ATEs with the ground-truth coefficient, we point out that all the methodologies are quite accurate as per mean values. Full matching shows the highest standard deviation (one order higher than the others), while OLS regression has the smallest standard deviation. This is indeed expected as the relationship between *weight* and *treat* is actually linear.

To conclude, it is worth noticing that on average OLS regression is the fastest, as it requires just 0.005 s to run. Exact Matching and IPW are also pretty fast, taking around 0.05 s to compute the parameters. On the other hand, AIPW takes around 30 s to estimate the parameter, while Full Matching requires almost 1 min on each run.

5 | CURRENT TRENDS AND POTENTIAL FUTURE RESEARCH

In recent years, the demand for trustworthy AI systems fostered the introduction of causality approaches in machine learning (ML) research. As highlighted by Pearl (2018), causal reasoning is crucial to overcome current ML limitations. For instance, the widespread usage of black-box models to socially sensitive decision making requires explainations of the logic involved (Guidotti et al., 2019). In fact, traditional ML algorithms build on the correlation among variables rather than on proper causal structures, with the risk of making wrong, biased, or harmful decisions.

In the eXplainable AI (XAI) branch, several works started exploiting causal frameworks to investigate black boxes decisions (Moraffah et al., 2020). CXPlain system, developed by Schwab and Karlen (2019), exploits Granger-causality to determine the importance of features for a black box model. Causal concepts, such as counterfactuals, are used in XAI (Verma et al., 2020) for post-hoc explanations which answer the question "which changes in an instance's features would have changed the ML model prediction?" Early approaches generate counterfactuals by solving an optimization problem (Wachter et al., 2017). A number of refinements of the optimization constraints cover efficiency, robustness, diversity, actionability, and plausibility. Zhao and Hastie (2021) exploit the connection between partial dependence plots, an XAI tool, and the backdoor criterion, to extract causal information from black-box models. Beyond post-hoc explanations, causal discovery algorithms can provide inherently interpretable methods (Xu et al., 2020).

Causal reasoning in AI also supports the enforcement of ethical concerns. In fact, spurious correlations and other forms of bias can lead to discriminatory decisions again protected-by-law social groups. Fair ML aims at the design of models that do not discriminate. The approach of (Kusner et al., 2017) defines fair decisions using counterfactual reasoning. Counterfactual fairness addresses what would have happened if membership to the protected group had been different and the other features had been the same, by exploiting the use of DAGs and do-operations. For a comprehensive review of fairness and causality, see Makhlouf et al. (2020).

Another concern about ML algorithms is the assumption that training and test set data are from the same distribution. In real environments, such an assumption is often not met due, for instance, to distribution shifts (Quiñonero-Candela et al., 2009). Domain adaptation studies how to extend ML models that are trained in certain domains to others. This branch can benefit from the usage of causal tools, as going beyond simple correlations is crucial to achieve robustness. Research bridging causality with domain adaptation includes Zhang et al. (2015), where the authors study which knowledge is transferable from one domain to another and find optimal target-domain hypothesis. In Pearl and Bareinboim (2014), the authors present conditions for transferring the causal effects learned in experimental studies to a new population where only observational studies are possible. Such a problem is identified as transportability. Transportability requires full knowledge of the causal graph. An extension to cases where there is partial knowledge of the causal structure is considered in Magliacane et al. (2018), where the authors build on JCI (Section 3.1).

Reinforcement learning (RL) is another branch of AI that can benefit from causal reasoning. In RL, an agent interacts with an environment aiming to maximize its cumulative reward within a certain time horizon.

Causality is a natural addition to model-based RL, where the agent has access to the state-transition probabilities, as it allows to overcome some concerns, such as confounding factors. For instance, Bareinboim et al. (2015) explore the relationship between causal models with unobserved confounders and the popular sequential decision-making setting of Multi-Armed Bandits (MAB), where latent variables influence both the reward distribution and the agent's intuition. Similar results for Markov Decision Processes (MDP), another popular setting for modeling sequential decision making, are provided by Zhang and Bareinboim (2016). In Lee and Bareinboim (2018), the authors study how to identify the best action in MABs, where an action corresponds to interventions on an arbitrary causal graph, taking into account also latent factors. Unconfoundedness is also studied in Bruns-Smith (2021), where cases of persistent confounding variables are investigated. Another work worth mentioning for causal RL is Lattimore et al. (2016), where non-interventional observations are used to improve the rate at which high-reward actions can be identified.

Let us turn now on the challenges posed by the usage of causality in AI and on a few open research directions.

First, there is, in general, a lack of knowledge regarding the causal model that underlies the data generating process in the majority of the application context. This is a general concern, which also limits the validation of causal discovery techniques, as few datasets are documented with a proper causal structure. This is a main challenge to be addressed, which requires a multi-disciplinary effort to collect domain expertise. Legal initiatives which demand for trustworthy AI development, such as the European Union draft regulation on AI,[10] can be a stimulous for the development of domain knowledge in the form of a causal graph in specific high-risk application domains.

Second, causal discovery research needs a boost to deal with non-tabular data, such as images and videos. Some examples of works in this area can be found in Lopez-Paz et al. (2017) and Li et al. (2020). The former work deals with retrieving causal signals in the context of images, while the latter studies the problem of learning causal structures from videos. In this context, a repository with publicly available datasets and tools is currently missing, and it would be a cornerstone for research in the area. The study of techniques designed for longitudinal data is also worth exploring, as few methodologies are available to retrieve causal structures (Section 3.2). Dealing with high-dimensional data and missing values is another critical open issue worth being addressed.

Third, some lines for future research can be outlined for causal inference. The assumption of unconfoundedness is often too strong, which makes inapplicable the tools surveyed in Section 4.1.2. Additionally, the overlap condition rarely holds in contexts with high-dimensional data. Also, the SUTVA assumption can quickly fail when network effects take place. Finally, methodological guidelines should be developed for specific application contexts to inform and guide practitioners in using the growing range of causal inference techniques.

6 | CONCLUSIONS

Causality is a dynamic and multidisciplinary field with both a long history and large developments to come. Several new techniques have been proposed in the last decades, and new software has become available for practitioners to perform causality-related tasks.

In light of this, this survey paper recollected together the principal methodologies, tools, datasets, and metrics to perform and evaluate both causal discovery and causal inference. Current trends of using causality in the realm of (trustworthy) Artificial Intelligence was provided, including open issues and potential new directions. A companion website (https://tinyurl.com/Causal-Discovery-and-Inference) is also available with additional resource lists and software scripts.

ACKNOWLEDGEMENTS

Open Access Funding provided by Scuola Normale Superiore within the CRUI-CARE Agreement. [Correction added on 25 May 2022, after first online publication: CRUI funding statement has been added.]

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Ana Rita Nogueira: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Salvatore Ruggieri:**

Conceptualization (equal); data curation (equal); funding acquisition (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Dino Pedreschi:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Joao Gama:** Conceptualization (equal); data curation (equal); formal analysis (equal); funding acquisition (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); software (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in “Methods and Tools for Causal Discovery and Causal Inference” at https://github.com/AnaRitaNogueira/Methods-and-Tools-for-Causal-Discovery-and-Causal-Inference.git.

ORCID

Ana Rita Nogueira https://orcid.org/0000-0002-2005-1943
Andrea Pugnana https://orcid.org/0000-0001-9138-8212
João Gama https://orcid.org/0000-0003-3357-1195

ENDNOTES

[1] https://tinyurl.com/Causal-Discovery-and-Inference.

[2] A skeleton is a graph with only undirected edges.

[3] More examples can be found in the companion website.

[4] https://bnlearn.com/.

[5] https://cran.r-project.org/web/packages/pcalg/index.html, and in the Causality Workbench.

[6] http://www.causality.inf.ethz.ch.

[7] https://www.bnlearn.com/documentation/man/asia.html.

[8] More examples can be found in the companion website.

[9] http://math.furman.edu/dcs/courses/math47/R/library/lmtest/html/ChickEgg.html.

[10] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, *42*(13), 1–17.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, *93*(1), 113–132.

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781–807.

Aliprantis, D. (2015). A distinction between causal effects in structural and Rubin causal models. *SSRN Electronic Journal*, *15-05*.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2019). *Synthetic difference in differences [Technical report]*. National Bureau of Economic Research.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, *5*(2), 37–51.

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

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2016). The great time series classification bake-off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660.

Bareinboim, E., Forney, A., & Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada* (pp. 1342–1350). NIPS.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. *Evaluation Studies*, *5*, 43–59.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, *3*(1), 1–31.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Knowledge discovery in databases workshop* (pp. 359–370). AAAI Press.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2014). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, *9*, 247–274.

Brooks-Gunn, J., Liaw, F.-R., & Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics*, *120*(3), 350–359.

Bruns-Smith, D. (2021). Model-free and model-based policy evaluation when causality is uncertain. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research* (pp. 1116–1126). PMLR.

Bühlmann, P., Kalisch, M., & Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: Partially faithful distributions and the pc-simple algorithm. *Biometrika*, *97*(2), 261–278.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2021). *Rdrobust: Robust data-driven statistical inference in regression-discontinuity designs. R package version 1.0.2*. R Foundation for Statistical Computing.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772.

Caron, A., Manolopoulou, I., & Baio, G. (2020). Estimating individual treatment effects using non-parametric regression models: A review. *CoRR* abs/2009.06472.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2021). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press.

Chatton, A., Borgne, F. L., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., & Foucher, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*, *10*(1), 9219.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). *Causalml: Python package for causal machine learning*. CoRR.

Chen, L., & Ng, R. (2004). On the marriage of lp-norms and edit distance. In *Proceedings 2004 VLDB Conference* (pp. 792–803). Elsevier.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research*, *2*, 445–498.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, *15*(1), 3741–3782.

Colombo, D., Maathuis, M. H., Kalisch, M., & Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, *40*(1), 294–321.

Comandé, G. (2018). Opinions - the rotting meat error: From Galileo to Aristotle in data mining? *European Data Protection Law Review*, *4*(3), 270–277.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43.

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

Dhir, A., & Lee, C. M. (2020). Integrating overlapping datasets using bivariate causal discovery. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020* (Vol. 34, pp. 3781–3790). AAAI Press.

Dieng, A., Liu, Y., Roy, S., Rudin, C., & Volfovsky, A. (2019). Interpretable almost-exact matching for causal inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2019)* (Vol. 89, pp. 2445–2453). PMLR.

Dimmery, D. (2016). *Rdd: Regression discontinuity estimation. R package version 0.57*. R Foundation for Statistical Computing.

Entner, D., & Hoyer, P. O. (2010). On causal discovery from time series data using fci. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models* (pp. 121–128). Helsinki Institute for Information Technology (HIIT).

Fong, C., Ratkovic, M., & Imai, K. (2021). *CBPS: Covariate balancing propensity score. R package version 0.22*. R Foundation for Statistical Computing.

Foster, D. J., & Syrgkanis, V. (2020). Statistical learning with a nuisance component (extended abstract). In C. Bessiere (Ed.), *International Joint Conference on Artificial Intelligence (IJCAI 2020)* (pp. 4726–4729). IJCAI ijcai.org

Foucher, Y., Le Borgne, F., Dantan, E., Gillaizeau, F., Chatton, A., & Combescure, C. (2020). *RISCA: Causal Inference and Prediction in Cohort-Based Analyses. R package version 0.9*. R Foundation for Statistical Computing.

Fox, J., Kleiber, C., & Zeileis, A. (2021). *ivreg: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics. R package version 0.6-0*. R Foundation for Statistical Computing.

Gerhardus, A., & Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in neural information processing (NeurIPS 2020)*. NIPS.

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

Glynn, A., & Quinn, K. (2017). *CausalGAM: Estimation of Causal Effects with Generalized Additive Models. R package version 0.1-4*. R Foundation for Statistical Computing.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*(1), 36–56.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424.

Gruber, S., & van der Laan, M. J. (2012). Tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*(13), 1–35. https://doi.org/10.18637/jss.v051.i13

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93:1–93:42.

Guo, R., Cheng, L., Li, J., Hahn, P. R., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, *53*(4), 75:1–75:37.

Guyon, I., Aliferis, C., Cooper, G., Elisseeff, A., Pellet, J. P., Spirtes, P., & Statnikov, A. (2011). Causality workbench. In *Causality in the sciences*. Oxford University Press.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*, 315–331.

Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, *13*, 2409–2464.

Hausser, J., & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*, 1469–1484.

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

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Hitchcock, C. (2020). Causal Models. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Hmamouche, Y. (2020). Nlints: An R package for causality detection in time series. *The R Journal*, *12*(1), 21.

Ho, D. E., Imai, K., King, G., Stuart, E. A., et al. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Huang, Y., & Kleinberg, S. (2015). Fast and accurate causal inference from time series data. In *International Florida Artificial Intelligence Research Society Conference (FLAIRS 2015)* (pp. 49–54). AAAI Press.

Iglesias, F., & Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, *6*(2), 579–597.

Imai, K., & Ratkovic, M. (2013). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *76*(1), 243–263.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29.

Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science*, *29*(3), 323–358.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jabbari, F., Ramsey, J. D., Spirtes, P., & Cooper, G. F. (2017). Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. In *European conference on machine learning and knowledge discovery in databases (ECML PKDD 2017), Vol. 10535 of LNCS* (pp. 142–157). Springer.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(11), 1–26.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. Springer-Verlag.

Kleinberg, S. (2012). *Causality, probability, and time*. Cambridge University Press.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4156–4165.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems (NIPS 2017)* (pp. 4066–4076). NIPS.

Lattimore, F., Lattimore, T., & Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain* (pp. 1181–1189). NIPS.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B: Methodological*, *50*(2), 157–194.

Lechner, M. (1999). Earnings and employment effects of continuous gff-the-job training in East Germany after unification. *Journal of Business & Economic Statistics*, *17*(1), 74–90.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165–224.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, *142*(2), 675–697.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lee, S., & Bareinboim, E. (2018). Structural causal bandits: Where to intervene? In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 2573–2583). NIPS.

Li, Y., Torralba, A., Anandkumar, A., Fox, D., & Garg, A. (2020). Causal discovery in physical systems from videos. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, Virtual*. NIPS.

Lopez-Paz, D., Nishihara, R., Chintala, S., Schölkopf, B., & Bottou, L. (2017). Discovering causal signals in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017* (pp. 58–66). IEEE Computer Society.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., & Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 10869–10879). NIPS.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions. *CoRR* abs/2010.09553.

Mao, H., & Li, L. (2018). *PSW: Propensity Score Weighting Methods for Dichotomous Treatments. R package version 1.1-3*. R Foundation for Statistical Computing.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. American Psychological Association.

Microsoft Research. (2019). EconML: A Python Package for ML-based heterogeneous treatment effects estimation. Version 0.x.

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, *21*, 99:1–99:108.

Moraffah, R., Karami, M., Guo, R., Raglin, A., & Liu, H. (2020). Causal interpretability for machine learning—Problems, methods and evaluation. *SIGKDD Explorations*, *22*(1), 18–33.

Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., Raglin, A., & Liu, H. (2021). Causal inference for time series analysis: Problems, methods and evaluation. *CoRR*, *63*, 3041–3085.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Nogueira, A. R., Gama, J., & Ferreira, C. A. (2021). Causal discovery in machine learning: Theories and applications. *Journal of Dynamics & Games*, *8*, 203–231.

Ogarrio, J. M., Spirtes, P., & Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In A. Antonucci, G. Corani, & C. P. de Campos (Eds.), *International Conference on Probabilistic Graphical Models (PGM 2016)* (Vol. 52, pp. 368–379). JMLR.

Orlandi, V., Roy, S., Rudin, C., & Volfovsky, A. (2020). *FLAME: Interpretable Matching for Causal Inference. R package version 2.0.0*. R Foundation for Statistical Computing.

Papke, L. E. (2005). The effects of spending on test pass rates: Evidence from Michigan. *Journal of Public Economics*, *89*(5–6), 821–839.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the Conference of the Cognitive Science Society*, pp. 329–334.

Pearl, J. (1989). *Probabilistic reasoning in intelligent systems—Networks of plausible inference. Morgan Kaufmann series in representation and reasoning*. Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. In Y. Chang, C. Zhai, Y. Liu, & Y. Maarek (Eds.), *Proceedings of the eleventh ACM international conference on web search and data mining, WSDM 2018, Marina Del Rey, CA, USA, February 5–9, 2018* (p. 3). ACM.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579–595.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Petersen, M. L., Wang, Y., Van Der Laan, M. J., & Bangsberg, D. R. (2006). Assessing the effectiveness of antiretroviral adherence interventions: Using marginal structural models to replicate the findings of randomized controlled trials. *Journal of Acquired Immune Deficiency Syndromes*, *43*, S96–S103.

Plato, H. (1961). *Phaedo* (pp. 40–98). Princeton University Press.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., & Schwaighofer, A. (2009). *Dataset shift in machine learning*. The MIT Press.

Raghu, V. K., Poon, A., & Benos, P. V. (2018). Evaluation of causal structure learning methods on mixed data types. In *ACM SIGKDD Workshop on Causal Discovery, CD@KDD 2019* (Vol. 92, pp. 48–65). PMLR.

Ramanan, N., & Natarajan, S. (2020). Causal learning from predictive modeling for observational data. *Frontiers in Big Data*, *3*, 535976.

Ramsey, J. D., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, *3*(2), 121–129.

Ramsey, J. D., Zhang, K., Glymour, M., Romero, R. S., Huang, B., Ebert-Uphoff, I., Samarasinghe, S., Barnes, E. A., & Glymour, C. (2018). Tetrad—A toolbox for causal discovery. In *8th International Workshop on Climate Informatics*.

Reiersøl, O. (1945). *Confluence analysis by means of instrumental sets of variables (PhD thesis)*. Almqvist & Wiksell.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9–12), 1393–1512.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Ropponen, O. (2011). Reconciling the evidence of Card and Krueger (1994) and Neumark and Wascher (2000). *Journal of Applied Econometrics*, *26*(6), 1051–1057.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Runge, J. (2004–2021). Tigramite—Causal discovery for time series datasets. Retrieved from https://tocsy.pik-potsdam.de/tigramite.php/.

Runge, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International conference on artificial intelligence and statistics (AISTATS 2018)* (Vol. 84, pp. 938–947). PMLR.

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In R. P. Adams & V. Gogate (Eds.), *Conference on Uncertainty in Artificial Intelligence (UAI 2020)* (Vol. 124, pp. 1388–1397). AUAI Press.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996.

Sachs, K. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523–529.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634.

Schwab, P., & Karlen, W. (2019). Cxplain: Causal explanations for model interpretation under uncertainty. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada* (pp. 10220–10230). NIPS.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn r package. *Journal of Statistical Software*, *35*(3), 1–22.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, *42*(7), 1–52.

Shalit, U., Johansson, F. D., & Sontag, D. A. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *International conference on machine learning (ICML 2017)* (Vol. 70, pp. 3076–3085). PMLR.

Sharma, A., Kiciman, E., et al. (2019). *DoWhy: A python package for causal inference*. Microsoft Research.

Sheppard, K., Lewis, B., Guangyi, Wilson, K., Thrasibule, Rene-Corail, X., & Vikjam. (2021). *Bashtage/linearmodels: Release 4.24 (Version 4.24)*. Zenodo.

Singh, K., Gupta, G., Tewari, V., & Shroff, G. (2018). Comparative benchmarking of causal discovery algorithms. In S. Ranu, N. Ganguly, R. Ramakrishnan, S. Sarawagi, & S. Roy (Eds.), *ACM India joint international conference on data science and Management of Data (COMAD/CODS 2018)* (pp. 46–56). ACM.

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, *97*(1), 1–66.

Spirtes, P. (2001). An anytime algorithm for causal inference. In T. S. Richardson & T. S. Jaakkola (Eds.), *Proceedings of the eighth international workshop on artificial intelligence and statistics, AISTATS 2001, Key West, Florida, USA, January 4–7, 2001*. Society for Artificial Intelligence and Statistics.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search, 2nd ed. Adaptive computation and machine learning*. MIT Press.

Spirtes, P., Meek, C., & Richardson, T. S. (1995). Causal inference in the presence of latent variables and selection bias. In P. Besnard & S. Hanks (Eds.), *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18–20, 1995* (pp. 499–506). Morgan Kaufmann.

Stigler, M., & Quast, B. (2016). *Rddtools: A toolbox for regression discontinuity in R*. Maison de la Paix.

Strobl, E. V. (2019). Improved causal discovery from longitudinal data using a mixture of dags. In *ACM SIGKDD workshop on causal discovery (CD@KDD 2019)* (Vol. 104, pp. 100–133). PMLR.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Theil, H. (1961). *Economic forecasts and policy*. North-Holland.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317.

Thurman, W. N., & Fisher, M. E. (1988). Chickens, eggs, and causality, or which came first? *American Journal of Agricultural Economics*, *70*(2), 237–238.

Tibshirani, J., Athey, S., & Wager, S. (2020). *Grf: Generalized random forests. R package version 1.2.0*. R Foundation for Statistical Computing.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78.

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1), 1–38.

van der Wal, W. M., & Geskus, R. B. (2011). Ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, *43*(13), 1–23.

Verma, S., Dickerson, J. P., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *CoRR* abs/2010.10596.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*, 841–887.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wang, T., Morucci, M., Awan, M. U., Liu, Y., Roy, S., Rudin, C., & Volfovsky, A. (2021). FLAME: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, *22*, 31:1–31:41.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage Learning.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. Macmillan Company.

Xu, G., Duong, T. D., Li, Q., Liu, S., & Wang, X. (2020). Causality learning: A new perspective for interpretable machine learning. *CoRR* abs/2006.16789.

Yagoubi, D. E., Akbarinia, R., Kolev, B., Levchenko, O., Masseglia, F., Valduriez, P., & Shasha, D. E. (2018). Parcorr: Efficient parallel methods to identify similar time series pairs across sliding windows. *Data Mining and Knowledge Discovery*, *32*(5), 1481–1507.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, *15*(5), 74:1–74:46.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems (NeurIPS 2018)* (pp. 2638–2648). NIPS.

Yu, K., Liu, L., & Li, J. (2021). A unified view of causal and non-causal feature selection. *ACM Transactions on Knowledge Discovery from Data*, *15*(4), 63:1–63:46.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.

Zhang, J., & Bareinboim, E. (2016). *Markov decision processes with unobserved confounders: A causal approach (Technical report)*. Purdue AI Laboratory.

Zhang, K., Gong, M., & Schölkopf, B. (2015). Multi-source domain adaptation: A causal view. In B. Bonet & S. Koenig (Eds.), *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA* (pp. 3150–3157). AAAI Press.

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, *39*(1), 272–281.

Zhou, T., Tong, G., Li, F., Thomas, L., & Li, F. (2021). *PSweight: Propensity score weighting for causal inference with observational studies and randomized trials. R package version 1.1.4*. R Foundation for Statistical Computing.

How to cite this article: Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., & Gama, J. (2022). Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1449. https://doi.org/10.1002/widm.1449