International Journal of Data Science and Analytics (2019) 8:285–297
https://doi.org/10.1007/s41060-018-0168-0

REGULAR PAPER

Data-driven discovery of causal interactions

Saisai Ma[1] · Lin Liu[1] · Jiuyong Li[1] · Thuc Duy Le[1]

Received: 24 March 2018 / Accepted: 20 December 2018 / Published online: 12 January 2019

Abstract
Causal discovery is a primary focus in many fields. Various methods have been developed to mine causal relationships from observational data. Most of the methods are only capable of identifying individual causes without considering their interactions. However, in real life, many effects are due to multiple factors that interact with each other. Therefore, detecting the interactions between those causal factors is essential for understanding the real causal mechanisms. So far, there are no efficient data-driven approaches to discovering causal interactions from data, especially large data sets. In this paper, we propose a general data-driven framework and develop four algorithms instantiated from the framework to detect causal interactions, directly from data. Extensive experiments on both synthetic and real-world data have shown that the proposed framework and the algorithms can achieve high effectiveness and efficiency for causal interaction discovery.

Keywords Causal discovery · Potential outcome · Causal interactions

1 Introduction

Everything occurs with reasons. An immature death may be caused by the malnutrition in childhood, a lack of exercise in teens, smoking in youth, bad diets in middle age, and a family history of heart attacks, etc. Often it is the case that not only the causal factors alone, but also the interactions between the factors lead to an immature death.

The study of the interactions between multiple causal factors (called *causal interactions* hereafter) is indeed very useful, as the knowledge of causal interactions has many real-world applications [3,37]. For example, it has been increasingly accepted that many diseases are resulted from not only genetic defects and environmental exposures, respectively, but also the interactions between them [8]. Knowing such interactions is helpful for understanding and preventing diseases.

In this paper, we study causal interactions and develop an efficient method to detect them directly from data. The interaction refers to the case that risk in the exposure to multiple factors simultaneously cannot be explained by the individual effects of these factors. In other words, the combined effect of multiple variables (e.g. a genetic factor G and an environmental factor E) on an outcome Y is different from the additive effects of multiple variables considered separately (i.e. the addition of G's effect and E's effect). For instance, genes for skin pigmentation (e.g. MC1R) and high-level sunlight exposure each have effect on getting skin cancer, but the risk of having skin cancer is much higher when both factors appear at the same time [8]. An interaction is causal, only if the interaction always exists in all conditions (when covariates having different values).

It is essential to differentiate causal interactions from additive effects of multiple variables, to determine whether or not these variables can be studied separately without losing important characteristics resulting from the interactions [34]. For example, it is biased (even not correct) to separately study the effect of asbestos exposure on lung cancer, if investigators ignore the interaction between asbestos exposure and smoking status [18].

The concept of causal interactions is also different from the following types of causal relationships involving multiple factors:

- multiple causes [23], which focus on all the individual causes of a variable that are represented by a causal Bayesian network or a local causal structure;
- conditional causal relationship [19,28], which concerns the relationship between a cause and the outcome, under a specific condition; and

✉ Saisai Ma
saisai.ma@mymail.unisa.edu.au

[1] School of IT and Mathematical Sciences, University of South Australia, Adelaide, Australia

– combined causes [20], which consider multiple variables as a single combined variable and examine the effect of the combined variable on the outcome, without distinguishing the individual effects from the effect due to variable interactions.

In contrast, the study on causal interaction is to assess the interaction between multiple individual variables that promotes to produce or prevent an outcome.

Causal interactions have attracted many attentions from domain experts, in clinical study, epidemiology, psychology, and etc. Various methods have been developed to study this problem [22,30]. However, these methods are hypothesis driven, i.e. they are used for validating hypothesised causal interactions. The hypothesis about a causal interaction has to be established beforehand, which requires prior knowledge and is often difficult to achieve, especially when there are a large number of possible interactions.

On the other hand, in many application areas, we have an abundance of data available. There are data mining or machine learning methods for finding interactions between multiple variables [6,21] from data directly, but the interactions found are normally association based and they may not be causal.

Rothman [29] developed the sufficient-component cause model to define a natural and logical view of causation. Under the model, the occurrence of a causal mechanism (also known as a sufficient cause), comprising a set of component causes, inevitably results in the occurrence of the outcome, and the component causes of a same sufficient cause always have causal interactions to lead to the outcome. Van der Weele and Robins [35] have made progress in detecting causal interactions based on this model. Although they have introduced some conditions for detecting causal interactions, the work largely stays at theoretical level and no specific algorithms have been developed for the detection.

Jiang et al. [10] developed the multiple beam search algorithm (MBS) to identify interacting genes associated with a disease from data, by learning local causal relationships. The MBS algorithm firstly does a greedy forward search and adds the predictor (gene) in each iteration that increases the Bayesian score the most. Then a greedy backward search is performed to get the minimal gene set. And they claim that the genes in the set have causal interactions with each other. However, if any predictor in the gene set has a strong individual effect, they will typically be scored highly and be considered as an interaction even if they do not interact.

To the best of our knowledge, there is no concrete computational methods for discovering the causal interactions between variables, especially from large data sets. In this paper, we aim to develop efficient data-driven methods for causal interaction discovery, by bringing together the principle of the well-established potential outcome model and efficient data mining approaches. The contributions of this paper are summarised as follows:

1. We study the problem of causal interaction discovery from a data mining perspective and elaborate the computational challenges for such discoveries.
2. We present a general framework, the data-driven approach to causal interaction discovery (DACID) and develop multiple instantiations of the framework to discover causal interactions. They are the first data mining algorithms for discovering causal interactions between multiple variables.
3. The experiments with both synthetic and real-world data sets are performed to demonstrate the effectiveness and efficiency of the proposed algorithms.

In the rest of this paper, the problem definition is presented in Sect. 2. In Sect. 3, we formally define the concept of causal interactions and develop the conditions to detect causal interactions. The DACID framework is proposed to mine causal interactions from observational data in Sect. 4. Section 5 demonstrates the effectiveness and efficiency of the proposed algorithms by experiments. Section 6 reviews the related work. Finally, we conclude the paper in Sect. 7.

2 Problem definition

In this section, we define the research problem studied in this paper. Before defining the problem, we firstly differentiate two concepts: statistical interactions and causal interactions, and then introduce the potential outcome model [31], the cornerstone for building the methods to detect causal interactions.

2.1 Notation

We use upper and lower case letters, e.g. X and x, to represent a random variable and its value, respectively. Bold-faced upper and lower case letters, e.g. $\boldsymbol{X}$ and $\boldsymbol{x}$, represent a set of variables and the corresponding values, respectively. We use the symbol "\" to denote the set difference operator, and we use the shorthand, e.g. $\boldsymbol{X}_{\setminus i}$ to represent $\boldsymbol{X} \setminus \{X_i\}$. Particularly, we denote the set of predictor variables and the target variable with $\boldsymbol{V}$ and Y, respectively.

In this paper, the predictor variables that are to be tested for possible causal interactions and the target variable are required to be binary, i.e. having two possible values, 1 or 0. The presence of a binary variable X means that the value of X is equal to 1. For a binary set of variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$, $\boldsymbol{X} = \mathbf{1}$ denotes that $X_i = 1$ ($i = \{1, \ldots, m\}$) and $Y_{\boldsymbol{X}=\mathbf{1}}$ represents the value of Y when $\boldsymbol{X} = \mathbf{1}$, where $\mathbf{1}$ means a

unit vector. We also use $\boldsymbol{x} \leq \boldsymbol{x'}$ to represent $x_i \leq x'_i$ for all $i \in \{1, \ldots, m\}$.

2.2 Statistical and causal interactions

In statistics, an interaction is typically defined based on the departure from additive effects of multiple variables on an outcome [13]. In other words, there exists an interaction between two variables G and E; the effect when G and E appearing together is different from the additive effect of the two variables, when they appear separately.

A natural way for assessing the statistical interaction between two variables is to measure the difference between the combined effect of the variables and the individual effect of each variable [30]. Let p_{ij} denote the risk of a specific disease based on the status of the genetic factor G and environmental factor E, $i, j \in \{0, 1\}$. Thus, the interaction between the two factors can be measured by the *risk difference*:

$$(p_{11} - p_{00}) - [(p_{10} - p_{00}) + (p_{01} - p_{00})], \tag{1}$$

where $(p_{11} - p_{00})$ represents the effect of both factors together compared to the reference category (i.e. both factors are absent) and $(p_{10} - p_{00})$ and $(p_{01} - p_{00})$ denotes the effects of the genetic and environmental factors, respectively. Note that Eq. (1) is exchangeable with $[(p_{11} - p_{01}) - (p_{10} - p_{00})]$ and $[(p_{11} - p_{10}) - (p_{01} - p_{00})]$, which are interpreted as the difference between the effects of one factor when another factor has different values.

If the result of Eq. (1) is not equal to zero, it is said that there exists an interaction between genetic and environmental factors on the disease. Instead of using *risk differences*, one may use *risk ratios*, *Odds Ratios*, or *relative excess risk due to interaction* to measure the statistical interactions [30].

Most existing methods only measure the association-based relationships that are not necessary to be causal, as they did not take the effects of confounding variables into account. However, to assess a causal interaction, it is essential to control potential covariates (confounders) for adjustment [1]. We have been aware of some interaction detection methods considering confounding elimination, but they typically work under the assumption of the absence of confounding [37].

Vanderweele et al. [35] developed a theory to detect the presence of causal interactions based on the sufficient-component cause model. With the theory, causal interactions between multiple predictor variables are examined under each stratum defined by the confounding variables, and there exists a causal interaction between predictor variables only if the interaction appears in all strata. However, instead of providing a concrete exploration approach, the theory can only be applied to validate hypothesised causal interactions, where it is required to generate hypotheses about predictor variables potentially having interactions and the confounding variables beforehand based on domain knowledge.

Compared with these existing approaches, we aim to develop a computational method for discovering the causal interactions between multiple variables directly from data, where no domain knowledge is required. To this end, we take advantages of data mining techniques for computational efficiency and a well-established causal model, the potential outcome model for causal examination.

2.3 The potential outcome model

The potential outcome model [31] is a major framework for causal inference, specifically for estimating causal effects in observational or experimental studies.

Let X denote a predictor variable (or treatment variable) and Y be the target variable. The individual ω receiving the treatment (i.e. $X_\omega = 1$) is in the treatment group, while the one not receiving the treatment (i.e. $X_\omega = 0$) is in the control group. In the potential outcome model, for individual ω, the causal effect of the treatment is the difference between the outcomes of Y if ω receiving the treatment, Y^1_ω, and receiving the control, Y^0_ω, i.e. $\delta_\omega = Y^1_\omega - Y^0_\omega$.

We often aggregate the causal effects of individuals to obtain the average causal effect (ACE), $E[\delta_\omega]$, i.e. the expected value of causal effects for all individuals, as follows:

$$E[\delta_\omega] = E[Y^1_\omega] - E[Y^0_\omega] \tag{2}$$

In fact, we cannot observe both Y^1_ω and Y^0_ω at the same time, because only Y^1_ω can be observed if individual ω receives the treatment, and vice versa. However, in an ideal (purely randomised) study, ACE can be estimated as the difference in the average outcomes between the treatment and control groups, i.e.

$$E^{ideal}[\delta_\omega] = E[Y^1_\omega | X_\omega = 1] - E[Y^0_\omega | X_\omega = 0] \tag{3}$$

In observational studies, it is not possible to randomly assign a treatment, and the covariates (denoted as $\boldsymbol{C}$ in this paper) make difference between individuals in a data set, which produces bias and affects the estimation of ACE. A perfect stratification on the covariates stratifies the data into a number of strata, such that all individuals in each stratum are indistinguishable, except the state of the treatment. Thus, we can estimate the ACE within each stratum $\boldsymbol{C} = \boldsymbol{c}$ as:

$$\begin{aligned} \text{ACE}_{\boldsymbol{c}} &= E^{ideal}[\delta_\omega | \boldsymbol{C} = \boldsymbol{c}] \\ &= E[Y^1_\omega | X_\omega = 1, \boldsymbol{C} = \boldsymbol{c}] - E[Y^0_\omega | X_\omega = 0, \boldsymbol{C} = \boldsymbol{c}] \end{aligned} \tag{4}$$

The ACE in a population can be determined by aggregating the ACEs in the strata:

$$\mathrm{ACE} = \sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}}\, \mathrm{ACE}_{\boldsymbol{c}} \tag{5}$$

where $w_{\boldsymbol{c}}$ stands for the weight of the stratum $\boldsymbol{c}$.

2.4 Problem definition

In this paper, we develop the practical definitions and detection criteria of causal interactions, based on the potential outcome model. We also take the advantages of data mining techniques to develop an efficient framework to detect causal interactions from observational data.

Normally there exist two types of causal interactions [30]: (1) positive causal interactions, also named superadditive interactions in [34], which facilitate to produce the outcome, and (2) negative causal interactions, also named subadditive interactions in [34], which are to prevent the occurrence of the outcome. Both kinds of causal interactions are to strengthen the overall effect of multiple variables on the target. The only difference is that positive causal interactions increase the effects in the positive direction (i.e. producing the outcome), while negative causal interactions increase the effects in the negative direction (i.e. preventing the outcome).

For example, if doctors suggest us take two or more drugs, it is expected that these drugs will have causal interactions and together help us to recover from the disease more effectively. That is, the interactions positively increase the effect of the drugs on recovery. An example of negative causal interaction is that both high price and high maintenance cost have negative effects on the action of buying a car, respectively, while the negative causal interaction between these two variables will strengthen their individual negative effects, and thus, the presence of both variables results in a lower chance of buying a car.

To define the casual interactions, we begin with the definition of a monotonic effect, which allows for the construction of powerful statistical tests [35].

Definition 1 (Monotonic Effect [35]) Let $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ and Y be a set of binary variables and a binary target, respectively. If we have $E[Y|\boldsymbol{X} = \boldsymbol{x}] \geq E[Y|\boldsymbol{X} = \boldsymbol{x'}]$ whenever $\boldsymbol{x} \geq \boldsymbol{x'}$, then $\boldsymbol{X}$ has positive monotonic effects on Y. If we have $E[Y|\boldsymbol{X} = \boldsymbol{x}] \leq E[Y|\boldsymbol{X} = \boldsymbol{x'}]$ whenever $\boldsymbol{x} \geq \boldsymbol{x'}$, then $\boldsymbol{X}$ has negative monotonic effects on Y.

In this paper, we make an assumption that the effects of treatment variables are monotonic, as [12,35] did. For the simplicity of the presentation, we only introduce the definition of positive causal interactions under the positive monotonic effect assumption and the corresponding criteria to detecting positive causal interactions. They can be easily adapted under the assumption of negative monotonic effects to define and detect negative causal interactions by changing only the direction of the effects of variables. In the experiments, we will show the results of the discovery of both positive and negative causal interactions from data.

Definition 2 (Positive Causal Interactions) Let a variable set $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ satisfy the assumption of positive monotonic effects. If $\exists X_i \in \boldsymbol{X}$ s.t. $\mathrm{ACE}_{\{X_i, \boldsymbol{X}_{\setminus i}=\boldsymbol{1}\}} - \sum_{j\neq i} \mathrm{ACE}_{\{X_i, X_j=0, \boldsymbol{X}_{\setminus\{i,j\}}=\boldsymbol{1}\}} > 0$, i.e. the causal effect of X_i with the presence (participation) of all remaining $m-1$ variables exceeds the sum of the causal effects of X_i with the presence of $m-2$ (i.e. 1 less) variables, then $\boldsymbol{X}$ has m-way positive causal interactions.

In Definition 2, we use $\mathrm{ACE}_{\{X_i, \boldsymbol{X}_{\setminus i}=\boldsymbol{1}\}}$ to denote the average causal effect of X_i with all variables in $\boldsymbol{X}$ except X_i set to 1, and $\mathrm{ACE}_{\{X_i, X_j=0, \boldsymbol{X}_{\setminus\{i,j\}}=\boldsymbol{1}\}}$ represents the average causal effect of X_i with $X_j = 0$ and $\boldsymbol{X}_{\setminus\{i,j\}} = \boldsymbol{1}$.

The research problem in this work is to detect causal interactions from data, which is shown as follows.

Problem 1 Given an observational data set $\boldsymbol{D}$ for the predictor variables $\boldsymbol{V}$ and the target Y, find the variable sets $\boldsymbol{X} \subseteq \boldsymbol{V}$, s.t. $\boldsymbol{X}$ has causal interactions with respect to the target Y by Definition 2.

3 Detecting causal interactions

In this section, we present the criterion for detecting causal interactions. We firstly illustrate the main idea for detecting the 2-way causal interactions, the interaction involving 2 variables, then we introduce the criterion for detecting m-way ($m \leq 2$) causal interactions.

3.1 2-Way causal interactions

Taking the study of the causes of lung cancer as an example, let $X_1 = 1$ or 0 denote *occupational exposure to asbestos* or not, $X_2 = 1$ or 0 denote *smoking* or not and $Y_{x_1x_2}$ represent *suffering the lung cancer* or not when $X_1 = x_1$ and $X_2 = x_2$, $x_1, x_2 \in \{0, 1\}$. Then from Eq. (2), the average causal effect of X_1 when X_2 is present ($X_2 = 1$) and not present ($X_2 = 0$) can be, respectively, estimated as $\mathrm{ACE}_{\{X_1 1\}} = E[Y_{11}] - E[Y_{01}]$ and $\mathrm{ACE}_{\{X_1 0\}} = E[Y_{10}] - E[Y_{00}]$, where $\mathrm{ACE}_{\{X_1x_2\}}$ denotes the ACE of X_1 when $X_2 = x_2$. If we have $\mathrm{ACE}_{\{X_1 1\}} > \mathrm{ACE}_{\{X_1 0\}}$, i.e. the causal effect of X_1 with X_2's presence is larger than the causal effect of X_1 without X_2's presence, then based on Definition 2, we can conclude that there is a positive causal interaction between X_1 (*exposure to asbestos*) and X_2 (*smoking*) on Y (*lung cancer*). Note that the condition also can be presented as $\mathrm{ACE}_{\{1X_2\}} > \mathrm{ACE}_{\{0X_2\}}$.

The above conditions can be presented as the following criterion for testing the presence of a positive causal interaction between two variables.

Criterion 1 *Let X_1 and X_2 have positive monotonic effects on the target Y. There is a positive causal interaction between X_1 and X_2, if any of the following conditions is met*

$$\mathrm{ACE}_{\{X_1 1\}} - \mathrm{ACE}_{\{X_1 0\}} > 0 \tag{6}$$

$$\mathrm{ACE}_{\{1 X_2\}} - \mathrm{ACE}_{\{0 X_2\}} > 0 \tag{7}$$

where $\mathrm{ACE}_{\{X_1 x_2\}}$ represents the causal effect of X_1 on Y when $X_2 = x_2$ and $\mathrm{ACE}_{\{x_1 X_2\}}$ the causal effect of X_2 on Y when $X_1 = x_1$.

Note that these two conditions (Eqs. (6) and (7)) are identical, if the ACE in the above conditions is expanded by the following Eq. (4).

Criterion 1 presents the conditions for detecting 2-way causal interactions, where we suppose there are no confounders, i.e. no other variables influencing the causal effects of X_1 and X_2 on the target Y. However, as mentioned in Sect. 2.3, randomised assignment of treatments is not possible in observational studies, so we employ the stratification strategy to address the confounding issue, which stratifies the data into a number of strata. Then we estimate the average causal effects in each stratum and aggregate them over all the strata (see Eqs. (4) and (5)). So we have the following criterion for detecting the 2-way positive causal interaction.

Criterion 2 *Let X_1 and X_2 have positive monotonic effects on the target Y and $\boldsymbol{C}$ be the covariates. There is a positive causal interaction between X_1 and X_2, if any of the following conditions is met*

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}} (\mathrm{ACE}_{\{X_1 1|\boldsymbol{c}\}} - \mathrm{ACE}_{\{X_1 0|\boldsymbol{c}\}}) > \alpha \tag{8}$$

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}} (\mathrm{ACE}_{\{1 X_2|\boldsymbol{c}\}} - \mathrm{ACE}_{\{0 X_2|\boldsymbol{c}\}}) > \alpha \tag{9}$$

where $w_{\boldsymbol{c}}$ is the weight of stratum $\boldsymbol{C} = \boldsymbol{c}$ and α is a threshold specified by users to filter the causal interactions with low strength.

Note that in this paper we assume that the differences of individuals could be captured by the covariates, i.e. the set of variables used for stratification. This assumption implies that there are no hidden confounding variables to bias the causal effect estimation.

3.2 *m*-Way causal interactions

Similar to 2-way causal interactions, we can derive the criterion for detecting m-way causal interactions from data. However, when $m > 2$, the situation becomes more complicated, since there exist more ways for selecting variables

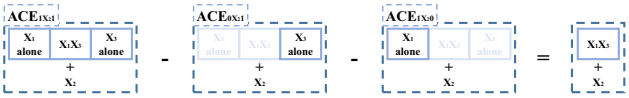


Fig. 1 An illustration of 3-way causal interactions

X_i and X_j (see Definition 2). In the following, we discuss how to do the variable selection for the m-way causal interactions. We firstly look at the case of 3-way (i.e. $m = 3$) causal interactions.

Given three binary variables, X_1, X_2, and X_3, if we focus on X_1, and consider the causal effect of X_1 when both X_2 and X_3 are present, in comparison with the causal effects when either X_2 or X_3 is absent, based on Definition 2, we have: if $\mathrm{ACE}_{\{X_1 11\}} - (\mathrm{ACE}_{\{X_1 01\}} + \mathrm{ACE}_{\{X_1 10\}}) > 0$, the three variables have a positive causal interaction. Similarly when we focus on X_2 and X_3, respectively, we obtain the other two conditions for 3-way causal interactions: $\mathrm{ACE}_{\{1 X_2 1\}} - (\mathrm{ACE}_{\{0 X_2 1\}} + \mathrm{ACE}_{\{1 X_2 0\}}) > 0$, and $\mathrm{ACE}_{\{11 X_3\}} - (\mathrm{ACE}_{\{01 X_3\}} + \mathrm{ACE}_{\{10 X_3\}}) > 0$. If any of these three conditions or criteria holds, we say that X_1, X_2, and X_3 have 3-way positive causal interactions with respect to the target Y.

To understand the physical meaning of the conditions for 3-way causal interactions, as illustrated in Fig. 1 (when focusing on X_2), we can consider that the overall causal effect of X_2 with the participation of X_1 and X_3 ($\mathrm{ACE}_{\{1 X_2 1\}}$) on Y in fact attributes to three elements: the causal effect of X_2 with the presence of X_1's contribution, the causal effect of X_2 with the presence of X_3's contribution, and the causal effect of X_2 with the presence of the combined contribution of X_1 and X_3. The 3-way causal interaction is indeed the third element, the causal effect of X_2 with the presence of the combined contribution of X_1 and X_3. Therefore, we take away $\mathrm{ACE}_{\{1 X_2 0\}}$ and $\mathrm{ACE}_{\{0 X_2 1\}}$ from the overall causal effect of X_2 with the participation of X_1 and X_3 ($\mathrm{ACE}_{\{1 X_2 1\}}$) to get the causal effect of the 3-way causal interaction.

Now let us develop the criterion for detecting m-way causal interactions. Firstly, we need to introduce the concept of a subordinate set [35], which enables us use the ACE concept and notation to present the causal effects of $m - 1$ variables on Y in the criterion to be developed (Criterion 3).

Definition 3 (Subordinate Set) Given a set of binary variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$, let $\boldsymbol{u}_m^i = \{(x_1, \ldots, x_m) \in \{0,1\}^m : \boldsymbol{x}_{\setminus i} = \boldsymbol{1}, x_i = 0\}$ and $\boldsymbol{u}_m^{i,j_i} = \{(x_1, \ldots, x_m) \in \{0,1\}^m : \boldsymbol{x}_{\setminus\{i,j_i\}} = \boldsymbol{1}, x_i = x_{j_i} = 0, i \neq j_i\}$. The set $\mathcal{S}_k = \{\boldsymbol{u}_m^{1,j_1}, \ldots, \boldsymbol{u}_m^{k-1,j_{k-1}}, \boldsymbol{u}_m^{k+1,j_{k+1}}, \ldots, \boldsymbol{u}_m^{m,j_m}\}$ is a subordinate set of order m, if $\exists l \in \{1, \ldots, k-1, k+1, \ldots, m\}$, s.t. $\boldsymbol{u}_m^{l,j_l} \preceq \boldsymbol{u}_m^k$.

In Definition 3, $\boldsymbol{u}_m^i$ ($\boldsymbol{u}_m^{i,j_i}$) represents an m-length vector with i-th (i-th and j_i-th) element(s) equal to 0 and the

remaining equal to 1. Now we use the following example to explain the subordinate set and the detection of the 3-way causal interaction.

Example 1 For the estimation of 3-way causal interactions ($m = 3$), if we focus on the causal effect of X_2 (i.e. $k = 2$) on Y, then $\mathcal{S}_2 = \{\boldsymbol{u}_3^{1,j_1}, \boldsymbol{u}_3^{3,j_3}\}$, which has three possible combinations based on Definition 3: $\{\boldsymbol{u}_3^{1,3}, \boldsymbol{u}_3^{3,2}\}$, $\{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,2}\}$ and $\{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,1}\}$. Any one of the three combinations can be selected for detecting 3-way causal interactions, e.g. $\mathcal{S}_2 = \{\boldsymbol{u}_3^{1,2}, \boldsymbol{u}_3^{3,2}\}$ (i.e. $i = 1$, $j_i = 2$ and $i = 3$, $j_i = 2$), and thus, a condition of 3-way causal interaction is expressed as $\mathrm{ACE}_{\{1X_21\}} - (\mathrm{ACE}_{\{0X_21\}} + \mathrm{ACE}_{\{1X_20\}}) > 0$.

As illustrated in the above example, the subordinate set $\mathcal{S}_k$ normally is not unique. The different conditions can be created based on different subordinate sets, and the causal interactions will exist, once any one condition is satisfied [35].

The formal criterion for detecting an m-way causal interaction is shown in the following, with the confounding taken into account.

Criterion 3 *$\boldsymbol{X} = \{X_1, \ldots, X_m\}$ satisfy the positive monotonic effect assumption, and $\boldsymbol{C}$ be the covariate set. The variable set $\boldsymbol{X}$ exhibits an m-way positive causal interaction, if $\exists X_k \in \boldsymbol{X}$ and some subordinate sets $\mathcal{S}_k$, s.t.*

$$\sum_{\boldsymbol{C}=\boldsymbol{c}} w_{\boldsymbol{c}} (\mathrm{ACE}_{\{X_k, \boldsymbol{X}_{\setminus k}=1|\boldsymbol{c}\}} - \sum_{\{i,j_i\}:\boldsymbol{u}_m^{i,j_i} \in \mathcal{S}_k} \mathrm{ACE}_{\{X_{j_i}, X_i=0, \boldsymbol{X}_{\setminus\{i,j_i\}}=\boldsymbol{1}|\boldsymbol{c}\}}) > \alpha,$$

where α is the threshold and $w_{\boldsymbol{c}}$ is the weight of the stratum $\boldsymbol{C} = \boldsymbol{c}$.

4 A data-driven approach to causal interaction discovery

In this section, we present the proposed framework, data-driven approach to causal interaction discovery (DACID) and the specific algorithms instantiated from the framework for discovering causal interactions in observational data. As shown in Framework 1, the algorithm includes three main steps: (1) candidate variable generation (lines 1–4), to only include the variables associated with the given target and to generate candidate sets for testing causal interactions; (2) data stratification (lines 10–11), to balance observed covariates between control and treatment groups to reduce bias; and (3) causal interaction discovery (lines 12–22), to detect the causal interactions between multiple variables with respect to the target. In each step, different data mining and statistical algorithms can be involved. Meanwhile, a pruning schema is employed to improve the efficiency of the algorithms.

FRAMEWORK 1: Data-driven Approach to Causal Interaction Discovery (DACID)

Input: A binary data set $\boldsymbol{D}$ for predictor variable set $\boldsymbol{V}$ and the target Y, the significant threshold α for testing causal interactions, and the maximum level of causal interactions k_0.
Output: $\boldsymbol{CI}_Y = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q\}$, where $\boldsymbol{X}_k \subseteq \boldsymbol{V}$ is a set of variables with causal interactions w.r.t Y.

1: $\boldsymbol{V}' \leftarrow predictorSelection(\boldsymbol{V}, Y)$
2: $\boldsymbol{CI}_Y \leftarrow \emptyset$
3: Let $m = 2$
4: Pairwise generate the 2-way candidate set $\boldsymbol{V}_2$ based on $\boldsymbol{V}'$
5: **while** $m \leq m_0$ **do**
6: **if** $m + 1 \leq m_0$ **then**
7: $(m+1)$-way candidate set $\boldsymbol{V}_{m+1} \leftarrow \emptyset$
8: **end if**
9: **for** each $\boldsymbol{X}_m$ in $\boldsymbol{V}_m$ **do**
10: $\boldsymbol{C} \leftarrow stratifyVariable(\boldsymbol{V}, \boldsymbol{X}_m, Y)$
11: $\boldsymbol{S} \leftarrow stratification(\boldsymbol{X}_m, \boldsymbol{D}, \boldsymbol{C})$
12: **for** each stratum $\boldsymbol{S} = \boldsymbol{s}$ **do**
13: $CIValue_{\boldsymbol{s}} \leftarrow causeInteraction(\boldsymbol{X}_m, Y, \boldsymbol{s})$
14: Get $w_{\boldsymbol{s}}$ from $\boldsymbol{s}$
15: **end for**
16: $CIValue \leftarrow \sum w_{\boldsymbol{s}} CIValue_{\boldsymbol{s}}$
17: **if** $CIValue > \alpha$ **then**
18: $\boldsymbol{CI}_Y \leftarrow \boldsymbol{CI}_Y \cup \boldsymbol{X}_m$
19: **if** $m + 1 \leq m_0$ **then**
20: for each $X \in \boldsymbol{V}' \setminus \boldsymbol{X}_m$, $\boldsymbol{V}_{m+1} \leftarrow \boldsymbol{V}_{m+1} \cup \{\boldsymbol{X}_m, X\}$
21: **end if**
22: **end if**
23: **end for**
24: $m = m + 1$
25: **end while**
26: Output $\boldsymbol{CI}_Y$

4.1 Candidate variable generation

If a set of variables $\boldsymbol{X}$ has a causal interaction with respect to a target, Y, then it is reasonable to assume that every variable in $\boldsymbol{X}$ is associated with Y. In this work, only variables associated with Y will be considered to generate the candidate sets for testing causal interactions. In Framework 1, the function $predictorSelection()$ in line 1 is used for the associated variable selection. Various methods can be implemented for this function as introduced in the following.

Correlation, Pearson correlation, Chi-square test are a commonly used criterion to describe the association between two variables. These methods for association analysis primarily vary in term of the types of variables (categorical or continuous variables). Under the proposed framework, one may use any of the association analysis methods suitable to their data types to select the associated variables. Once the associated variables selected, we pairwise generate the 2-way candidate set $\boldsymbol{V}_2$ (line 4) for testing causal interactions in the next sections.

4.2 Stratification

Stratification attempts to balance observed covariates by obtaining similar covariate distributions between treatment and control groups to reduce estimation bias. As described in Sect. 2.3, ideally, within a stratum where the distributions of covariates of the control group and the treatment group are indistinguishable, we can use Eq. (4) to obtain an unbiased estimation of the causal effects. Then we can aggregate the causal effects across all the strata. To this end, the key to the proper stratification of a data set is the choice of stratifying variables from the set of predictor variables.

In the DACID framework, we conduct stratification only on the covariates that are associated with the target (represented by stratifying variables $\boldsymbol{C}$). This is done by function *stratifyVariable*() in Framework 1 (line 10). For example, it is not necessary to control one person's facial features (e.g. a hawk nose), when estimating the causal interactions of asbestos exposure and the smoking status, regarding to lung cancer.

With the stratifying variable set selected, the simplest stratification is the perfect stratification (PS). PS requires that all individuals within a stratum have same values of the stratifying variables $\boldsymbol{C}$, to remove the effects of covariates on the target Y. Thus, PS is capable of eliminating bias in the estimations of causal effects and causal interactions. In the proposed DACID framework, *PS* is employed as an option when performing data stratification.

However, for a high-dimensional data set, *PS* may have a low statistical power for detecting dependency in data, as too many strata may be generated and each stratum has a small size. An alternative option, propensity score [26], is provided to increase the statistical power. The idea of propensity score is to stratify individuals to different strata, such that individuals in the same stratum have similar propensity scores. For an individual (sample), the propensity score is defined as the probability of the individual receiving the treatment T conditioning on the stratifying variables $\boldsymbol{C}$:

$$e(\boldsymbol{c}) = pr(T = 1|\boldsymbol{C} = \boldsymbol{c}).$$

However, normal propensity score method cannot handle our specific problem of detecting causal interactions, which involves multiple treatments. [9] proposed the generalised propensity score (GPS) to extend binary treatment to multiple treatments. The GPS is defined as the conditional probability of receiving a particular treatment t given $\boldsymbol{C}$:

$$g(t, \boldsymbol{c}) = pr(T = t|\boldsymbol{C} = \boldsymbol{c}).$$

Thus, each individual obtains a GPS vector $G(\boldsymbol{C}) = (g(t_1, \boldsymbol{c_1}), ..., g(t_Z, \boldsymbol{c_Z}))$, the conditional probabilities of receiving Z different treatments, respectively.

In the proposed framework, stratification on generalised propensity score (*SGPS*) is employed as an alternative option of (*PS*). We use multinomial logistic regression to obtain the GPS vector. K-means clustering is employed to separate individuals with similar generalised propensity score vectors into the same stratum. It has been shown that subclassification with 5 subgroups can remove at least 90% of the bias resulting from the covariates in the causal analysis [27,31]. And thus k is set to 5 in this work.

4.3 Causal interaction discovery

The stratification puts individuals with similar distribution of stratifying variables $\boldsymbol{C}$ in the same stratum. Within each stratum $\boldsymbol{S} = \boldsymbol{s}$, a contingency table is generated for the estimation of causal interactions. Now we illustrate the discovery of 2-way causal interactions.

Given a data set $\boldsymbol{D}$, X_1 and X_2 are binary predictors of the binary target Y. Each stratum $\boldsymbol{S} = \boldsymbol{s}$ of data $\boldsymbol{D}$ is divided into 4 treatment groups based on different values of X_1 and X_2 (see the contingency table below): *treatment* 1 ($\{X_1 = 1, X_2 = 1\}$), *treatment* 2 ($\{X_1 = 1, X_2 = 0\}$), *treatment* 3 ($\{X_1 = 0, X_2 = 1\}$), and *treatment* 4 ($\{X_1 = 0, X_2 = 0\}$), where n_{ij} ($i \in \{1, 2, 3, 4\}$, $j \in 1, 2$) denotes the frequencies of the values of X_1, X_2 and Y.

X_1	X_2	Y	
		1	0
1	1	n_{11}	n_{12}
1	0	n_{21}	n_{22}
0	1	n_{31}	n_{32}
0	0	n_{41}	n_{42}

Since individuals are indistinguishable in terms of the stratifying variables in the same stratum, no matter which treatment group the individuals are in, we can unbiasedly estimate the causal effects between any two treatment groups in the stratum. As an illustration, we use an example to show how to calculate the causal effect of X_2 in the cases of $X_1 = 1$ and $X_1 = 0$, respectively, and determine if there is a causal interaction between them.

Example 2 For clear presentation, we break the above contingency table into two 2×2 contingency tables, as shown in Table 1. Then the causal effect of X_2 in the case of $X_1 = 1$ can be estimated by comparing the individuals receiving *treatment* 1 ($\{X_1 = 1, X_2 = 1\}$) and *treatment* 2 ($\{X_1 = 1, X_2 = 0\}$), see the first contingency table in Table 1. Similarly, ACE between *treatment* 3 and *treatment* 4 groups is estimated from the second contingency table, to get the causal effect of X_2 when $X_1 = 0$. Probabilities (*Prob*) are the most common approach to calculating the ACEs [36].

Table 1 Two 2×2 contingency tables

X_1	X_2	Y		X_1	X_2	Y	
		1	0			1	0
1	1	n_{11}	n_{12}	0	1	n_{31}	n_{32}
1	0	n_{21}	n_{22}	0	0	n_{41}	n_{42}

$$\mathrm{ACE}_{\{1X_2|s\}} = n_{11}/(n_{11}+n_{12}) - n_{21}/(n_{21}+n_{22})$$
$$\mathrm{ACE}_{\{0X_2|s\}} = n_{31}/(n_{31}+n_{32}) - n_{41}/(n_{41}+n_{42})$$

Odds Ratio (*OR*) [5] is another widely used measure and may be more suitable to determine the ACEs when the response variable is binary [25].

$$\mathrm{ACE}_{\{1X_2|s\}} = n_{11}n_{22}/n_{12}n_{21}$$
$$\mathrm{ACE}_{\{0X_2|s\}} = n_{31}n_{42}/n_{32}n_{41}$$

With the calculations of *Prob* and *OR*, the causal effects of X_2 with $X_1 = 1$ and $X_1 = 0$ on the stratum $\boldsymbol{S} = \boldsymbol{s}$ are obtained. Then the causal interactions between X_1 and X_2 on Y can be determined based on Criterion 2, by aggregating ACEs in all strata. Here the weight w_s of one stratum $\boldsymbol{s}$ is set as the ratio of the sample size of $\boldsymbol{s}$ to the size of data $\boldsymbol{D}$.

The above is an example of 2-way causal interaction estimation. m-way causal interactions can be obtained with the similar progress.

4.4 Pruning schema and complexity analysis

Logically, if a set of k variables ($k \geq 3$) has a causal interaction on an outcome, then at least one of its subsets also has a causal interaction on the outcome. In other words, if a set of variables do not have causal interaction, any of its superset of variables does not have causal interaction either. Using this property, we can prune the search space by generating high-level causal interaction candidates based on confirmed lower-level interactions, as shown in line 19 of Framework 1. Thus, for the sake of the computational efficiency, the DACID framework identifies causal interactions in a level by level manner.

Now we analyse the time complexity of the proposed framework. In the case without the pruning schema, the number of all possible causal interaction candidate sets is $\sum_{k=2}^{|\boldsymbol{V}|}\binom{|\boldsymbol{V}|}{k}$, and the number of causal interaction tests is $O(2^{|\boldsymbol{V}|})$. For the case with the pruning schema, if the number of all k-level causal interaction candidate sets is $N(k)$ and the number of sets confirmed is $N'(k)$ ($N'(k) \ll N(k)$), then the number of $(k+1)$-level causal interaction candidate sets is $N(k+1) = N'(k)(|\boldsymbol{V}| - k)$. The total number of causal interaction candidate sets is $\sum_{k=2}^{|\boldsymbol{V}|-1} N'(k)(|\boldsymbol{V}| - k) + \binom{|\boldsymbol{V}|}{2}$, i.e. $O(|\boldsymbol{V}|^2)$. Therefore, the pruning schema significantly improves the efficiency of causal interaction discovery.

This pruning schema may bring about some false negatives, when the causal interaction discovery approach misses the lower-level causal interactions, since the higher-level (e.g. $(k+1)$-level) candidates are generated from the lower-level (e.g. k-level) causal interactions. Fortunately, for a $(k+1)$-level candidate, if any k-level causal interaction found is a subset of this $(k+1)$-level candidate, then this candidate will not be missed. For example, when we missed the causal interaction between X_2 and X_3, the candidate $X_1X_2X_3$ still could be generated if we detected the causal interaction between X_1 and X_2. Thus, the causal interaction discovery approach with the pruning schema can still obtain high-quality causal interaction discovery and high computational efficiency as well.

5 Experiments

In this work, Probability (*Prob*) and Odds Ratio (*OR*) are employed for the estimation of causal effects. Combined with two different stratification strategies, perfect stratification (*PS*) and stratification on generalised propensity score (*SGPS*), the proposed DACID framework is instantiated into four different algorithms: *PS-OR*, *PS-Prob*, *SGPS-OR*, and *SGPS-Prob*.

A collection of synthetic and real-world data sets are used to evaluate the effectiveness and efficiency of four instantiated algorithms of the proposed framework. We firstly run experiments on the synthetic data with known ground truth, where the threshold of detecting causal interactions is set to 0.15 (i.e. $\alpha = 0.15$ in Criterion 3 for all four algorithms), and compare the results with MBS [10] in Sect. 5.1. Then we run experiments on the real-world data sets, where 0.10 is set as the threshold to expect more potential causal interactions, and then, the results with strong causal interactions are extracted for detailed discussions (Sects. 5.2 and 5.3). Finally, we compare the efficiency between four instantiated algorithms with MBS on the synthetic datasets randomly generated by logistic regression (Sect. 5.4).

5.1 Synthetic data

To evaluate the proposed algorithms, we generate multiple binary synthetic data sets based on the definition of sufficient causes defined in [35]. According to the theorem in [29,30], there must exist causal interactions between component causes of each sufficient cause, and thus, the interactions among the component causes of a sufficient cause can be considered as the ground truth of causal interactions.

Specifically, three main steps are performed to generate a synthetic data set: (1) samples of the predictor variables

Table 2 Causal interactions discovered by the proposed algorithms and *MBS*

		V50-5K	V80-5K	V100-5K	V120-5K	V150-5K	V200-5K
		F_1 (SD)	F_1 (SD)	F_1 (SD)	F_1 (SD)	F_1 (SD)	F_1 (SD)
no noise	*PS-OR*	0.98 (0.05)	**1.00 (0.00)**	**1.00 (0.00)**	0.96 (0.08)	**1.00 (0.00)**	**1.00 (0.00)**
	PS-Prob	**1.00 (0.00)**	0.95 (0.06)	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**
	SGPS-OR	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	0.94 (0.08)	**1.00 (0.00)**
	SGPS-Prob	**1.00 (0.00)**	**1.00 (0.00)**	0.96 (0.06)	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**
	MBS	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	0.96 (0.07)	0.93 (0.11)
10% noise	*PS-OR*	**1.00 (0.00)**	0.95 (0.07)	**1.00 (0.00)**	**1.00 (0.00)**	0.94 (0.09)	0.93 (0.08)
	PS-Prob	**1.00 (0.00)**	**1.00 (0.00)**	0.96 (0.06)	**1.00 (0.00)**	0.95 (0.08)	**1.00 (0.00)**
	SGPS-OR	**1.00 (0.00)**	0.95 (0.07)	**1.00 (0.00)**	0.96 (0.06)	**1.00 (0.00)**	0.95 (0.07)
	SGPS-Prob	**1.00 (0.00)**	**1.00 (0.00)**	0.92 (0.08)	**1.00 (0.00)**	0.96 (0.07)	**1.00 (0.00)**
	MBS	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	0.98 (0.05)	0.92 (0.10)	0.90 (0.14)
20% noise	*PS-OR*	0.90 (0.09)	0.91 (0.10)	0.92 (0.09)	0.93 (0.08)	0.86 (0.08)	0.89 (0.07)
	PS-Prob	0.91 (0.09)	0.87 (0.08)	0.90 (0.08)	0.93 (0.08)	0.85 (0.07)	**0.91 (0.10)**
	SGPS-OR	**0.93 (0.07)**	0.89 (0.08)	0.92 (0.08)	**0.94 (0.08)**	0.87 (0.07)	0.89 (0.08)
	SGPS-Prob	0.90 (0.07)	0.89 (0.09)	**0.94 (0.09)**	0.91 (0.10)	**0.88 (0.09)**	0.89 (0.09)
	MBS	**0.93 (0.17)**	**0.95 (0.08)**	**0.94 (0.08)**	0.93 (0.10)	0.82 (0.27)	0.75 (0.15)

$\boldsymbol{V}$ and the target Y are randomly generated, (2) a subset of predictor variables $\boldsymbol{X}$ are randomly picked up from $\boldsymbol{V}$ and values of the variables in $\boldsymbol{X}$ are modified such that their values (and the values of Y) satisfy the definition of sufficient causes [35], i.e. let $\boldsymbol{X}$ be a sufficient cause of Y, and (3) repeat steps 2 and 3 to obtain another sufficient cause (and their samples) of Y. For example, a variable set $\boldsymbol{X} = \{X_1, X_2\}$ is picked up; then, we modify the samples of X_1 and X_2 such that for the whole data set we have $Y = 1$ once $X_1 = 1$ and $X_2 = 1$. In this way, there exists causal interaction between X_1 and X_2 based on the theorem in [30].

We have generated six data sets containing 50, 80, 100, 120, 150, and 200 variables, respectively, and each with 5K samples. To test the robustness of the proposed algorithms, we add random noise on 10% and 20% of samples for each predictor variable, respectively.

Table 2 shows the average results (F_1-measure) and standard deviations of 10 runs of experiments on the synthetic data sets. We can that see all of algorithms perform very well on the data sets without noise. The reason is that if two variables are sufficient to cause the occurrence of the target, then these two variables are more likely to have a strong causal interaction on the target, and thus, this type of causal interactions is easier to be discovered from the data.

As the noise increases, the performance of the proposed algorithms slightly decreases, but the accuracy still keeps above 0.85, while the *MBS* algorithm has a lower accuracy, especially on the data sets with more variables. Meanwhile, the standard deviations of the F_1-measure show that the proposed algorithms are more stable than *MBS*. Furthermore, *MBS* has a poor performance in terms of identifying the exact set of variables that have a real interaction with each other, as *MBS* is designed to greedily search a variable set, by adding variables into it so as to increase the overall causal effects. For example, the ground truth indicates that X_1 and X_2 have a causal interaction and X_3 does not interact with other variables, but *MBS* regards X_1, X_2, and X_3 have interactions, if they have a higher causal effect. Therefore, *MBS* fails to identify the exact pair with a real interaction, which in turn may generate a misleading understanding and action.

5.2 BRCA data

The BRCA (TCGA breast invasive carcinoma) contains the expression profiles of messenger RNAs (mRNAs) and microRNAs (miRNAs) of 753 cancer patients. MiRNAs are an important type of gene regulator, and there has been evidence that a group of miRNAs often co-regulate the same mRNAs, i.e. the miRNAs work together to cause the change of the expression levels of the same mRNAs [2]. As an example of demonstrating the performance of the proposed algorithms, we apply the algorithms to the BRCA data set to discover the causal interactions between the co-regulating miRNAs.

As our focus is on finding the causal interactions of the regulators (miRNAs), rather than finding miRNA targets (i.e. which miRNA regulates which mRNA), we pre-process the data set as follows. Firstly, we use the Limma package [24] to find the significantly differentially expressed miRNAs and mRNAs between the tumour and normal samples (p-value < 0.05, adjusted by the Benjamini–Hochberg (BH) method). The top 100 differentially expressed miRNAs are chosen as

Table 3 Statistical significance of validated miRNA–mRNA-miRNA causal interaction triplets

	BCL-2	RAS	VEGF
PS-OR	1.51E−16	7.81E−01	1.89E−03
PS-Prob	1.31E−12	8.39E−03	8.04E−03
SGPS-OR	1.10E−16	1.51E−02	1.09E−08
SGPS-Prob	1.46E−05	1.55E−01	5.07E−02

the predictor variables. Then we select 3 mRNAs (i.e. BCL-2, RAS and VEGF) from 5 differentially expressed mRNAs in the "pathways in cancer" as the targets. At the end, we obtain the data set containing the expression profiles of 100 miRNAs and 3 mRNAs.

We apply each of the four proposed algorithms to the pre-processed data set to discover 2-way causal interactions between the miRNAs. The predicted direct sequence binding information in TargetScan (v7.0) [15] is employed for the post-process, to filter the miRNAs that may be incapable of binding the 3 selected mRNAs. We denote an interaction output by our algorithms in the form of "miRNA–mRNA–mRNA, which represents (1) each miRNA is capable of binding the target mRNA, and (2) the two miRNAs have a cause interaction in co-regulating the mRNA.

The enrichment analysis is used to validate the quality of the relationships between miRNAs and mRNAs, by checking if (1) both of the interacting miRNAs and the target mRNA are in the same pathway, and (2) a miRNA pair having causal interactions has been experimentally confirmed to regulate the same target mRNA. Specifically, we focus on one significant KEGG (Kyoto Encyclopedia of Genes and Genomes) [11] pathway, "pathways in cancer". The MiRSEA package (https://cran.r-project.org/web/packages/MiRSEA/index.html) is used to retrieve the set of miRNAs that are in the pathway "pathways in cancer". The experimentally confirmed miRNA–mRNA regulatory relationships are downloaded from [14]. Based on the relationships, we identify the miRNA–mRNA–miRNA triplets.

A cumulative hypergeometric distribution model is then employed to assess the statistical significance of causal interactions discovered. Let N be the number of possible patterns (miRNA–mRNA–miRNA causal interaction triplets) in the pre-processed data set, K be the number of patterns in the data satisfying the above two conditions (i.e. in the same pathway and experimentally confirmed co-regulation relationships), n be the number of patterns discovered, and k be the number of validated patterns. The p-values of the validation results are obtained using the cumulative hypergeometric test formula:

$$p(X \geq k) = \sum_{i=k}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{10}$$

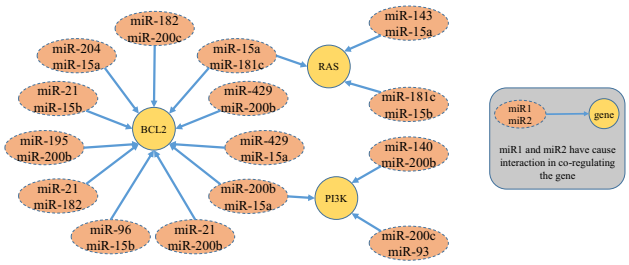


Fig. 2 Causal interactions discovered by at least three proposed algorithms, which are associated with the KEGG pathway, pathways in cancer, and have the experimentally confirmed co-regulation relationships

The top 50 detected causal interactions on each mRNA are extracted to analyse the performance of proposed algorithms. Table 3 shows the validated miRNA causal interactions on the three different mRNAs. The results show that the four proposed algorithms have a good performance in detecting causal interactions between miRNAs in co-regulating a target mRNA.

Figure 2 shows an example of the miRNA causal interactions discovered by at least three proposed methods with all miRNAs and mRNAs appearing in the "pathways in cancer", and the co-regulation relationships are experimentally confirmed.

5.3 METABRIC data

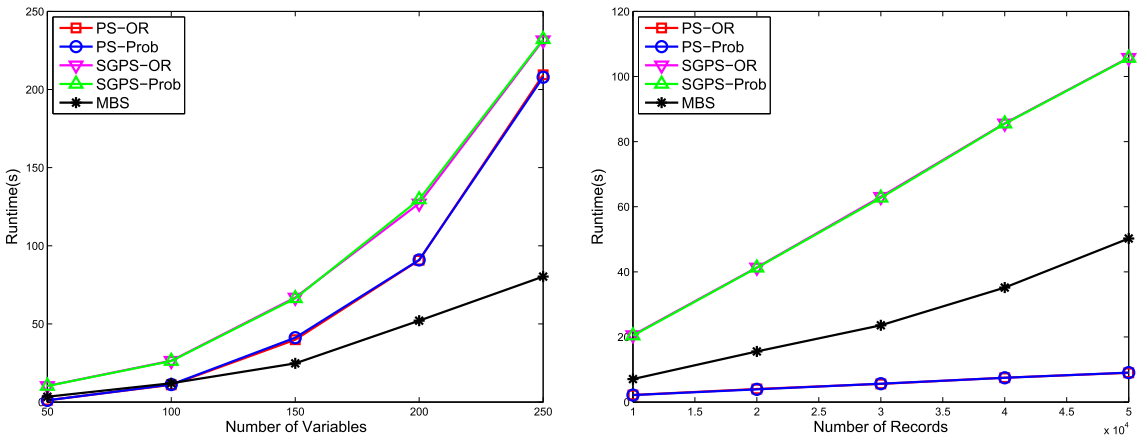
The METABRIC data set [33] contains clinical traits and outcomes for 1981 primary cancer tumours collected from participants of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) trial. For continuous variables, we convert them to categorical ones and then dichotomise them to get the binary data set. The transformed data set contains 35 binary variables and 1358 samples. The outcome variable in our experiments is 5-year survival, i.e. whether patients survive more than 5 years after being diagnosed as having breast cancer. In addition, the R package "impute" [7] is employed to impute missing data, when the maximum percentage of missing data in any row or column is less than 60%. If more than 60% of data is missed, then the row or column is dropped.

With the threshold setting of $\alpha = 0.10$, 42, 22, 63, and 19 causal interactions are detected by four instantiated algorithms. We rank the results based on the strength of causal interactions and extract TOP 5 positive and TOP 5 negative causal interactions for the evaluation and list the ones detected by at least two algorithms in Table 4. The results indicate that $chemo = no$ (the patient did not have chemotherapy) interacts with some features, e.g. $stage = 1$ (early stage), $size$ = 0-19 (small size of tumour), and $positive\_lymph = 0$ (no positive lymph nodes), to have a high chance to increase the 5-year survival rate. Chemother-

Table 4 Some causal interactions identified by four algorithms, from the METABRIC data set

Positive causal interactions (5-year survival)	*PS-OR*	*PS-Prob*	*SGPS-OR*	*SGPS-Prob*
$positive\_lymph = 0$ & $chemo = no$	√	√		√
$stage = 1$ & $chemo = no$	√	√		√
$diag\_age = 55$–69 & $chemo = no$	√		√	√
$size = 0$–19 & $chemo = no$	√	√		
$removed\_lymph = 4$–9 & $radiation = yes$			√	√
$ER.Expr = -$ & $radiation = yes$			√	√
negative causal interactions (5-year survival)	*PS-OR*	*PS-Prob*	*SGPS-OR*	*SGPS-Prob*
$ER.Expr = +$ & $chemo = yes$	√	√	√	√
$positive\_lymph = 2$–3 & $grade = 2$	√			√
$grade = 2$ & $chemo = yes$	√	√		
$diag\_age = 70$–84 & $positive\_lymph = 2$-3		√	√	
$stage = 3$ & $Er.Expr = +$		√		√
$stage = 3$ & $hormone = yes$			√	√
$positive\_lymph = 1$ & $removed\_lymph = 1$-3			√	√

diag_age age at diagnosis of the disease; *size* size of tumour in cm; *grade* grade of disease; *stage* composite of size and # positive nodes; *pos_lymph* # positive lymph nodes; *removed_lymph* # lymph nodes removed; *ER.Expr* oestrogen receptor expression; *chemo* whether patient had chemotherapy; *radiation* whether patient had radiation therapy; *hormone* whether patient had hormone therapy



(a) Scalability with the number of variables. **(b)** Scalability with the number of samples.

Fig. 3 Scalability evaluation

apy usually is not recommended for patients with non-invasive breast cancer, which may be a reason that $chemo = no$ interacts with various features to get positive results.

5.4 Scalability evaluation

In Sect. 4.4, we have analysed the computational complexity of the proposed framework. To experimentally assess its efficiency, we conduct scalability evaluation on the instantiated algorithms. We run *PS-OR*, *PS-Prob*, *SGPS-OR*, and *SGPS-Prob* on 10 synthetic data sets and compare the runtime with *MBS*. The data sets are randomly generated by using logistic regression, where the predictor variables and the target variable are binary. All the scalability evaluation experiments are run on the same computer with a 3.4 GHz Quad-core CPU and 16 GB of memory.

Figure 3a shows the running time of the algorithms using the data sets of the same sample size (5K) but different numbers of variables (50, 100, 150, 200 and 250). The results show that the five algorithms all scale well, but *MBS* is most efficient with respect to the number of variables. The algorithms using *PS* (*PS-OR* and *PS-Prob*) are slightly more efficient than the ones with *SGPS* (*SGPS-OR* and *SGPS-Prob*), since *PS* is faster than *SGPS*. Another observation is that the scalability of *PS-OR* and *PS-Prob* (*SGPS-OR* and *SGPS-Prob*) is the same. It is because that the different measures of the ACE estimation (*OR* or *Prob*) do not impact the efficiency of algorithms, when the stratification strategies are the same.

We then apply the algorithms to the data sets with 50 variables but different sample sizes (10K, 20K, 30K, 40K, and 50K). The execution time is shown in Fig. 3b. *PS-OR* and *PS-Prob* are much faster than other three methods consistently for different record sizes, while the running time of *SGPS-OR* and *SGPS-Prob* increases sharply as the data sets get larger. The main reason is that propensity scores are obtained using the logistic regressions, whose time complexity is polynomial to the number of samples. Thus, *SGPS-OR* and *SGPS-Prob* may not handle data sets with too large sample size, while *PS-OR*, *PS-Prob*, and *MBS* scale well with both number of variables and number of samples.

6 Related work

Numerous methods have been proposed to address the problem of causality, but most of them are designed to discover the causal relationship between a single factor and the outcome [4,23,32]. A little but growing literature seeks to detect multi-factor causes consisting of two or more component variables [16,17,20]. Although these methods were able to capture the causal relationships between multi-factor predictor variables and the target, they focus on measuring the combined effect of multiple variables on the outcome, instead of quantifying the interactions between these variables.

Novick et al. [22] developed theories to test conjunctive causes, which act in concert to produce or prevent an effect, and to detect the interactions between causes. Contingency information was used to judge interactions between two causal candidates [38]. These methods were designed to validate hypothesised causal interactions, which may be difficult to be generated even based on domain knowledge.

More recently, data mining and machine learning strategies are applied to learn interactions from large data sets. [6] proposed a maximum entropy probability model to search for genomic interactions on disease risks. [21] developed a multi-factor dimensionality reduction (MDR) method for collapsing high-dimensional genetic data into a single dimension and then detected interactions in relatively small sample sizes. By employing Bayesian network learning and information gain, [10] have developed a new method to discover interacting single nucleotide polymorphism (SNPs) and successfully detected some insights from real-world data. [39] have formalised the concept of synergistic interaction and applied it to causal inference.

In the past decade, Vanderweele et al. [35] have made progress in detecting causal interactions under the sufficient-component cause model [29]. However, the work largely stays at theoretical level and it is difficult to be applied to exploring causal interactions directly from data.

Distinct from these methods, the proposed DACID framework is capable of discovering the interactions between multiple individual variables, each of which may or may not be a cause of an outcome. Interactions identified by DACID are causal interactions w.r.t. the outcome, since the effects of covariates have been eliminated during the estimation of causal interactions. Moreover, DACID can be easily instantiated for the exploration of causal interaction from data, with no domain knowledge required.

7 Discussion and conclusion

Causal interaction discovery is an important topic in the field of causal discovery, as understanding the interactions between causal factors helps us gain valuable insights into the underlying causal mechanisms. It also provides us an alternative and effective way for identifying important causal factors.

The research of causal interactions has had a long history, but finding the interactions directly from data is still a new and challenging topic. There is a severe lack of data-driven approaches to discovering causal interactions, particularly from large data sets.

We have set up our ultimate goal to tackle the challenges by bringing together traditional causal discovery methods and efficient data mining techniques. This paper presents the outcome of our first and important step towards this goal.

In this paper, the concept of causal interactions is re-formalised by considering the detection of causal interactions as a data mining task, such that they can be applied to discovering causal interactions from observational data. A general framework has been developed to address the problem of causal interaction discovery around a given target variable. The framework is instantiated in various ways. The resulting algorithms are sound under the monotonic effect assumption.

In the set of experiments on synthetic data sets, all instantiated algorithms achieve a high accuracy for causal interaction discovery. The proposed algorithms have a good performance in detecting causal interactions between miRNAs in co-regulating a target mRNA. Meanwhile, experiments on the clinical data sets have shown that the proposed framework can find many causal interactions justifiable based on domain knowledge. The algorithms also achieve a high computational efficiency, especially for algorithms using perfect stratification.

In the near future, we will apply the proposed framework and algorithms to solve real-world problems, such as identifying the co-regulation mechanisms of multiple types (instead of just one type) of gene regulators, such as miRNAs and transcription factors.

Acknowledgements This work has been partially supported by Australian Research Council (ARC) Discovery grant DP140103617 and ARC Discovery grant DP170101306.

References

1. Ahrens, W., Krickeberg, K., Pigeot, I.: An introduction to epidemiology. In: Ahrens, W., Pigeot, I. (eds.) Handbook of Epidemiology, pp 1–40. Springer, Berlin (2005)
2. Bartel, D.P.: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116**(2), 281–297 (2004)
3. Dao, B., Nguyen, T., Venkatesh, S., Phung, D.: Latent sentiment topic modelling and nonparametric discovery of online mental health-related communities. Int. J. Data Sci. Anal. **4**(3), 209–31 (2017)
4. Eberhardt, F.: Introduction to the foundations of causal discovery. Int. J. Data Sci. Anal. **3**(2), 81–91 (2017)
5. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley, New York (2013)
6. Hahn, L.W., Ritchie, M.D., Moore, J.H.: Multifactor dimensionality reduction software for detecting gene–gene and gene-environment interactions. Bioinformatics **19**(3), 376–382 (2003)
7. Hastie, T., Tibshirani, R., Narasimhan, B., Chu, G.: Package 'impute' (2016). https://bioconductor.org/packages/release/bioc/manuals/impute/man/impute.pdf
8. Hunter, D.J.: Gene-environment interactions in human diseases. Nat. Rev. Genet. **6**(4), 287–298 (2005)
9. Imbens, G.W.: The role of the propensity score in estimating dose–response functions. Biometrika **87**(3), 706–710 (2000)
10. Jiang, X., Neapolitan, R.E., Barmada, M.M., Visweswaran, S., Cooper, G.F.: A fast algorithm for learning epistatic genomic relationships. AMIA Ann. Symp. Proc. **2010**, 341–345 (2010)
11. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1), 27–30 (2000)
12. Knol, M.J., VanderWeele, T.J., Groenwold, R.H.H., Klungel, O.H., Rovers, M.M., Grobbee, D.E.: Estimating measures of interaction on an additive scale for preventive exposures. Eur. J. Epidemiol. **26**(6), 433–438 (2011)
13. Kupper, L.L., Hogan, M.D.: Interaction in epidemiologic studies. Am. J. Epidemiol. **108**(6), 447–453 (1978)
14. Le, T.D., Zhang, J., Liu, L., Li, J.: Ensemble methods for miRNA target prediction from expression data. PLoS ONE **10**(6), e0131-627 (2015)
15. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., Burge, C.B.: Prediction of mammalian microRNA targets. Cell **115**(7), 787–798 (2003)
16. Li, J., Le, T.D., Liu, L., Liu, J., Jin, Z., Sun, B., Ma, S.: From observational studies to causal rule mining. ACM Trans. Intell. Syst. Technol. **7**(2), 14 (2015)
17. Li, J., Ma, S., Le, T., Liu, L., Liu, J.: Causal decision trees. IEEE Trans. Knowl. Data Eng. **PP**(99), 1–14 (2016)
18. Liddell, F.D.K.: The interaction of asbestos and smoking in lung cancer. Ann. Occup. Hyg. **45**(5), 341–356 (2001)
19. Ma, S., Li, J., Liu, L., Le, T.D.: Discovering Context Specific Causal Relationships. arXiv preprint arXiv:1808.06316 (2018)
20. Ma, S., Li, J., Liu, L., Le, T.D.: Mining combined causes in large data sets. Knowl. Based Syst. **92**, 104–111 (2016)
21. Miller, D.J., Zhang, Y., Yu, G., Liu, Y., Chen, L., Langefeld, C.D., Herrington, D., Wang, Y.: An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. Bioinformatics **25**(19), 2478–2485 (2009)
22. Novick, L.R., Cheng, P.W.: Assessing interactive causal influence. Psychol. Rev. **111**(2), 455 (2004)
23. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, Cambridge (2000)
24. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**(7), e47 (2015)
25. Robins, J.M.: Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, M.E., Berry, D. (eds.) Statistical Models in Epidemiology, the Environment, and Clinical Trials, pp 95–133. Springer, New York (2000)
26. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
27. Rosenbaum, P.R., Rubin, D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. **79**(387), 516–524 (1984)
28. Rosenblum, M., van der Laan, M.J.: Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. Biometrika **98**(4), 845–860 (2011)
29. Rothman, K.J.: Causes. Am. J. Epidemiol. **104**(6), 587–592 (1976)
30. Rothman, K.J., Greenland, S., Lash, T.L.: Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia (2008)
31. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688 (1974)
32. Song, J., Satoshi, O., Masahito, K.: Tell cause from effect: models and evaluation. Int. J. Data Sci. Anal. **4**(2), 99–112 (2017)
33. Soulakis, N.D., Carson, M.B., Lee, Y.J., Schneider, D.H., Skeehan, C.T., Scholtens, D.M.: Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J. Am. Med. Inf. Assoc. **22**(2), 299–311 (2015)
34. Van der Weele, T.J.: On the distinction between interaction and effect modification. Epidemiology **20**(6), 863–871 (2009)
35. Van der Weele, T.J., Robins, J.M.: A theory of sufficient cause interactions. COBRA Preprint Series, p. 13 (2006)
36. Van der Weele, T.J., Robins, J.M.: Empirical and counterfactual conditions for sufficient cause interactions. Biometrika **95**(1), 49–61 (2008)
37. Vimaleswaran, K.S., Power, C., Hyppnen, E.: Interaction between vitamin D receptor gene polymorphisms and 25-hydroxyvitamin D concentrations on metabolic and cardiovascular disease outcomes. Diabetes Metab. **40**(5), 386–389 (2014)
38. White, P.A.: Causal judgement from contingency information: judging interactions between two causal candidates. Q. J. Exp. Psychol. Sect. A **55**(3), 819–838 (2002)
39. Yang, S., Natarajan, S.: Knowledge intensive learning: combining qualitative constraints with causal independence for parameter learning in probabilistic models. In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, pp 580–595. Springer, Berlin (2013)