Causality Discovery with Domain Knowledge for Drug-Drug Interactions Discovery

Sitthichoke Subpaiboonkit$^{*[0000-0003-3503-4473]}$, Xue Li*,†, Xin Zhao$^{*[0000-0001-9121-6090]}$, Harrisen Scells$^{*[0000-0001-9578-7157]}$, and Guido Zuccon$^{*[0000-0003-0271-5563]}$

* The University of Queensland, Australia
{s.subpaiboonkit, x.zhao, h.scells, g.zuccon}@uq.edu.au
† Dalian Neusoft University of Information, China
lixue@neusoft.edu.cn

Abstract. Bayesian Network Probabilistic Graphs have recently been applied to the problem of discovery drug-drug interactions, i.e., the identification of drugs that, when consumed together, produce an unwanted side effect. These methods have the advantage of being explainable: the cause of the interaction is made explicit. However, they suffer from two intrinsic problems: (1) the high time-complexity for computing causation, i.e., exponential; and (2) the difficult identification of causality directions, i.e., it is difficult to identify in drug-drug interactions databases whether a drug causes an adverse effect – or vice versa, an adverse effect causes a drug consumption. While solutions for addressing the causality direction identification exist, e.g., the CARD method, these assume statistical independence between drug pairs considered for interaction: real data often does not satisfy this condition.
In this paper, we propose a novel causality discovery algorithm for drug-drug interactions that goes beyond these limitations: Domain-knowledge-driven Causality Discovery (DCD). In DCD, a knowledge base that contains known drug-side effect pairs is used to prime a greedy drug-drug interaction algorithm that detects the drugs that, when consumed together, cause a side effect. This algorithm resolves the drug-drug interaction discovery problem in $O(n^2)$ time and provides the causal direction of combined causes and their effect, without resorting to assuming statistical independence of drugs intake. Comprehensive experiments on real-world and synthetic datasets show the proposed method is more effective and efficient than current state-of-the-art solutions, while also addressing a number of drawbacks of current solutions, including the high time complexity, and the strong assumptions regarding real-world data that are often violated.

Keywords: Causality Discovery · Bayesian Network · Drug-Drug Interaction.

1 Introduction

An adverse effect is an undesired or harmful event caused by the consumption of a drug, or interactions between drugs. Adverse effects caused by any single drug

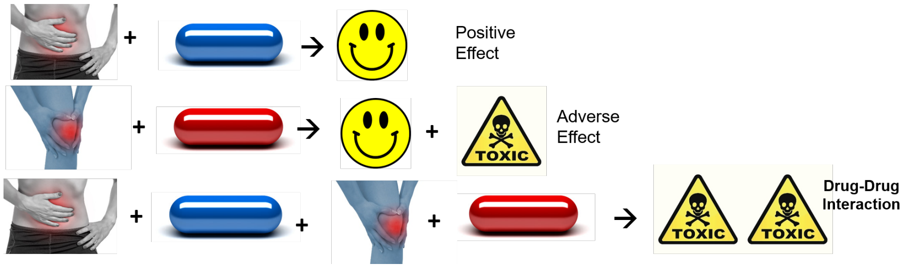

Fig. 1. An example of how two drugs, administered to treat two unrelated conditions, may interact to cause a severe adverse effect.

have been investigated in detail by medical researchers. However, it is often the case that an ill person would consume more than one drug at any given time, e.g., patients with AIDS or cancer usually need to consume a mixture of drugs at the same time [18]. In these cases, adverse effects can be caused by drug-drug interactions. A drug-drug interaction occurs when a consumed drug interacts with another consumed drug, e.g., aspirin consumed together with warfarin may cause excessive bleeding [5]; we denote this with the following notation $\texttt{aspirin} + \texttt{warfarin} \rightarrow \textit{\texttt{bleeding}}$. Adverse effects caused by drug-drug interactions are often more severe than those from single drugs. This is exemplified in Figure 1, where the blue drug effectively treats stomach ache with no side effect. The red drug treats knee pain, but it causes a non-severe adverse effect, with a certain likelihood. However, when taken together to treat stomachache and an unrelated knee pain, the likelihood of a severe adverse effect increases due to the drug-drug interaction.

Adverse effects caused by drug-drug interactions have a significant impact on public health. Adverse effects cause more than 100,000 deaths and 770,000 injuries per year in the United States alone, costing approximately USD 136.8 billion [9]; and around 30% of adverse effects are reported to be possibly caused by drug-drug interactions [17]. Adverse effects can be prevented if the cause of the drug-drug interaction is known. Unfortunately, it is difficult to perform biological experiments to discover the relation. These experiments in fact are costly, complicated and time-consuming. In addition, it is also impractical, if not impossible, to test all possible combinations of drugs via biological experiments, when more than two drugs are involved in a drug-drug interaction [2]. However, the availability of data related to suspected adverse effects as reported by health care authorities, health care providers, drug manufacturers and patients[1], offers the opportunity to study drug-drug interactions causing adverse effects (DDICAE) by using computational methods instead of biological experiments [3].

Most computational methods for discovery of DDICAE are based on statistical association or correlation [19, 16, 6]; however, they do not guarantee that the discovered relations between multiple drugs and adverse reactions are due

[1] For example, the Food and Drug Administration (FDA) in the United States has collected this type of data in the FDA Adverse Event Report System (FAERS).

to causal reasons. Previous attempts to model causal relations for the DDICAE problem, namely the Causal Association Rule Discovery (CARD) method [3], exist, but they are computationally infeasible, requiring exponential time to run when the number of drugs is considered for each reaction.

In this article, we propose a novel method for DDICAE discovery which extends upon the Bayesian Constraint-based Causality (BCC) model since BCC meaningfully represents causal relationships and effectively handles the large size of data sets, although it also has exponential running time. Our method aims to (1) solve the direction ambiguity problem[2] using conditional independence to prune causal and non-causal relations not involved in DDICAE and without resorting to the V-structure property of BCC[3], and (2) reduce the computational complexity of DDICAE discovery using greedy heuristics to select candidate drugs.

Our method exploits existing domain knowledge. For example, suppose the causal relation between a single drug and an adverse effect is already known (e.g., warfarin causes bleeding, $\texttt{warfarin} \rightarrow \textit{bleeding}$). Then, we can exploit this knowledge to identify which drug that, consumed together with the drug for which an adverse effect is known, increases the likelihood of the adverse effect to occur, e.g., aspirin consumed together with warfarin may increase the chances of excessive bleeding, $\texttt{aspirin} + \texttt{warfarin} \rightarrow \textit{bleeding} ++$. This knowledge is used to address the causal direction ambiguity problem because the known causal relation can be used to identify the causal direction within newly discovered causal relations. In addition, the domain knowledge can be exploited to reduce the computational complexity in combination with conditional independence by pruning candidate drugs that are unrelated to the interaction.

2 Related Works

In computational studies of DDICAE, correlations and associations are the key statistics being exploited, e.g., methods based on logistic regression [19], association rules [16] and bi-clustering [6]. These methods only focus on the correlation between drug-drug interactions and adverse effects to predict DDICAE, rather than finding the *causal* relations – which could provide superior insights into the relationship. Note in fact that correlation doesn't necessarily imply causation: causation happens when a change of the causal variable (e.g., consumption of both red and blue drugs) leads to a *direct* change of the effect variable (e.g., bleeding) [15]. In this work, causal variables are drugs, and effect variables are adverse effects. If causal relations between combined drug intake and adverse effect were known, then drug prescriptions could be adapted to prevent or mitigate the adverse effects.

Beyond methods that rely just on correlation or associations to discover DDICAE, computational methods that directly model causal relations have been

[2] The causal relationship whose direction is unknown

[3] The V-structure property, in fact, may not identify all causality structures in real-world DDICAE data [18], although it can identify the direction in causality discovery.

proposed. In Causal Bayesian Network (CBN), causes and effects are modelled using directed acyclic graphs; CBN has been shown to be effective beyond tasks in DDICAE [15, 20]. However, for practical purpose, CBN is computationally expensive when more than a few hundred variables are involved. In fact, in this method all possible network paths need to be considered and computation time grows exponentially based on the number of variables (this is an NP-Hard problem) [4]. A popular implementation of CBN is the PC algorithm [13].

The Bayesian Constraint-based Causality (BCC) model extends CBN to feasibly handle data with a large number of variables [1]. To obtain more efficient and scalable performance, BCC limits the discovery of causal relations only to local structures rather than considering the whole Bayesian graph as in CBN. However, in BCC, computation time depends on the number of combinations of variables in the conditional independence tests used to determine a causal relation – these can still be high, thus rendering the method impractical in real situations. In addition, most existing BCC-based approaches, such as Markov Blanket, can find causal relations between variables but cannot identify the direction of the relations (because they rely on measuring conditional independence for a local graph): we name this specific problem as *causality direction ambiguity*. Most BCC approaches can not solve the causality direction ambiguity problem.

Previous work has combined association rules with either partial association test [8] or cohort studies [10] to discover causal relations, including their direction. The partial association test is calculated after discovering association rules to confirm their causality – the key intuition of partial association test is similar to that of conditional independence. In cohort studies, causality relations are found using association rules on observational data fixing the specific control data. These methods are computationally impractical because their run time grows exponentially with respect to the size of variables in the data.

To our knowledge, Causal Association Rule Discovery (CARD) [3] is the only CBN-based method to discover causal relations along with their direction in DDICAE, thus comparable to the method we propose in this work. CARD uses association rules to select significant candidate drugs. Candidate drugs are then iteratively paired and tested for interactions [15] or common-effect relations (or V-structure)[4] [20]. However, this method makes restrictive hypotheses on the relations displayed by the data, and these are often not satisfied by real-world data. In addition, CARD becomes exponential with respect to the number of drugs, because it computes every possible combination – the method we propose, described in Section 3, instead offers a polynomial time ($O(n^2)$) solution to the problem of DDICAE discovery.

3 Proposed Method

The aim of DDICAE discovery is to identify two or more drugs that cause an adverse effect. Conventional DDI systems typically only consider two interacting

[4] The relationship describing the causes that are marginally independent become dependent when their common effect is given.

drugs. We introduce Domain-knowledge-driven Causality Discovery (DCD), a novel constraint-based method that uses conditional independence to discover causality. DCD is guided by exploiting pre-existing known drug-drug interaction adverse effect causation contained in domain knowledge resources. Our method is also made efficient by our novel approach to pruning unrelated candidate drugs.

Our method relies on conditional independence testing because an effective way to define causation between two variables is to (1) measure statistical dependence between the variables, and (2) ensure no other variables given as the condition eliminate their statistical dependence. Conditional independence can be exploited by iteratively determining whether to remove drugs that are not a direct cause of a target adverse effect. These steps are applied in the pruning stage of DCD to remove drugs that do not meet the necessary criteria to be considered candidate drugs (i.e., interacting drugs causing the target adverse effect). We define interacting drugs as conjunctive-combined-cause drugs. Finally, DCD employs a greedy optimisation step in order to find conjunctive-combined-cause drugs from candidate drugs that have the strongest statistical-dependence score.

Next we formally define the proposed method, and we detail the key steps of conditional independence testing and pruning; a compendium of the terminology and acronyms used in this paper is provided in Appendix I.

3.1 Formal Description of our Method

For the notation in this paper, we use $A \perp B$ to represent statistical independence between A and B. $A \perp B \mid C$ represents the conditional independence of A and B, given C. The causation between two or more variables is expressed by $\rightarrow$. The direction indicates which variables cause the other, e.g., $A \rightarrow C$, and $(A, B) \rightarrow C$ represents multiple causes of A and B causing C. We denote $D = \{d_1, d_2, ..., d_n\}$ as the drug consumption indicators for a patient, where d_i is the binary indicator representing whether a patient has consumed drug i (i.e., $d_i = 1$ if the patient uses this drug). Similarly, s denotes the binary occurrence indicator of an adverse effect. We further denote d_m as the drug from a domain knowledge resource that is known to cause an adverse effect s. Our model is provided as input a target adverse effect s, the domain knowledge drug d_m as a given cause, and other drugs D' excluding d_m (i.e., $D' = \left\{d_1, d_1, ..., d_n\right\}$ and $d_m \notin D'$). D' contains the possible drugs that form the conjunctive combined cause with d_m. Formally, our model identifies $O \subseteq D'$ drugs that are actually the conjunctive combined cause. Note that O can be empty if no drugs with significant interactions are found.

$$O = \underset{d_i \in D'}{\operatorname{argmax}} \; \texttt{Positive Dependence}(d_m, s \mid d_i) \tag{1}$$

Equation 1 describes how O is obtained. DCD selects O that has the largest dependence value (calculated using the Chi-squared test). Each candidate drug that is added to the combined-cause drugs set provides better positive PMI values than the conjunctive-combined-cause drugs set without that candidate

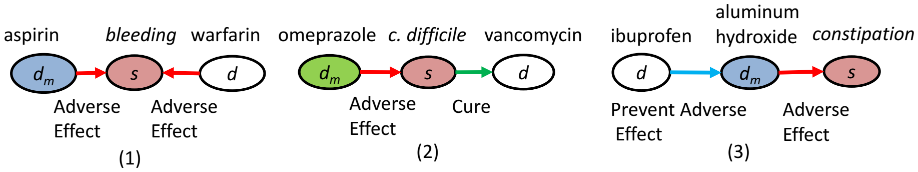

Fig. 2. Three examples of causal dependency (Sub-figures 1-3) that exist among the drugs: (1) domain knowledge drug and the adverse effect, (2) combined cause of s, and (3) consequence or cure of s.

drug. This approach selects the set of DDICAE that is most likely to be causes of the adverse effect.

3.2 Conditional Independence

Our method exploits the conditional independence and the domain knowledge drug and adverse effect to prune the unrelated drugs to the DDICAE (might be correlated or not). Then it uses a greedy algorithm based on Pointwise Mutual Information (PMI) and the Chi-squared statistical hypothesis test to select the highest positive dependence candidate drugs that have high probability to be DDICAE.

In DCD, correlation or dependence is important to screen for possible causal relations. We use the Chi-squared statistical hypothesis test [12] to test for independence. To find combined interacting drugs and the adverse effect, we focus primarily on the positive dependence relation as it accurately represents the co-occurrence DDICAE. We use both the Chi-squared test and Pointwise Mutual Information (PMI) [7] to measure positive dependence between $d_i \in D$ and the d_m which causes s. The Chi-squared test and PMI can solve the drawbacks of each other. The Chi-squared test finds the strength of the dependence when it is divided by the number of observations; however, it cannot discriminate between positive and negative dependence. In contrast, PMI cannot detect the strength of the dependence in sparse datasets where the number of samples with the same variables values is small. However, PMI can differentiate positive dependence from negative dependence.We thus use the Chi-squared test to find which $d_i \in D$ increases dependence strength of co-occurrence of the drug-drug interaction and s, and use PMI to select only positive dependence cases (e.g., $pmi(d_m = 1, d = 1; s = 1) > 0$).

3.3 Pruning

The pruning step aims to eliminate as many unrelated drugs as possible by identifying the drugs that are strongly correlated and positive dependent on d_m. This step also removes any drugs which have opposite causal direction with d_m that cause s. As the result, causal ambiguity (the causal relationship whose causal direction is unknown) is no longer a problem.

There are three possible cases (**C1**, **C2**, and **C3**) of causality existing among the drugs (d), the domain knowledge drug (d_m) and the adverse effect (s):

C1: $d_m \rightarrow s \leftarrow d$;
C2: $d_m \rightarrow s \rightarrow d$; and
C3: $d \rightarrow d_m \rightarrow s$

As illustrated in Figure 2.1, **C1** is the conjunctive combined cause (i.e., `aspirin` → *`bleeding`* ← `warfarin` indicates that both aspirin d_m and warfarin d causes bleeding s). In **C2** (Figure 2.2), the drug is the consequence of an adverse effect instead of a confounded (i.e., `omeprazole` → *`c. difficile`* → `vancomycin` indicates that omeprazole d_m and vancomycin d both cause c. difficile s). In **C3** (Figure 2.3), the candidate drug is the cause of d_m instead of the confounded of d_m (i.e., `ibuprofen` → `aluminum hydroxide` → *`constipation`* indicates that ibuprofen d is the reason to use aluminium hydroxide d_m, which is the cause of constipation s).

Equation 2 describes our pruning step. Here, the candidate drug d with the structure in **C2** and **C3** are removed from the candidate set D'. Any candidate drugs with any structures in **C2** are removed using conditional independence: $d_m \perp d \mid s = 1$. Any candidate drugs d with **C3** can also be removed using conditional independence: $d \perp s \mid d_m = 1$. In addition, drugs that are independent to d_m and s, or drugs that can be considered as unrelated to d_m and s are also removed when $(d_m \perp d \mid s = 1) \vee (d \perp s \mid d_m = 1)$. In this pruning process, the remaining candidate drugs have a higher probability to cause s. Next, the greedy process of our algorithm is performed to select the most suitable candidate drugs to be the combined causes.

$$D' = \begin{cases} D' - \{d\}, & \text{if } (d_m \perp d \mid s = 1) \ \vee \ (d \perp s \mid d_m = 1) \\ D', & \text{otherwise} \end{cases} \tag{2}$$

The pruning step continues each iteration in the greedy method. A candidate drug is removed if it is dependent on O using the conditional independence method as describes in Equation 3.

$$D' = \begin{cases} D' - \{d\}, & \text{if } (d \perp s \mid d_m, O = 1) \\ D', & \text{otherwise} \end{cases} \tag{3}$$

We next describe algorithmically our iterative pruning algorithm. In each iteration, based on the detected conjunctive combined cause from previous iterations, the drug which has the highest probability to cause s is considered as the candidate drug. If the candidate drug is a correct causal drug, then it is inserted into a conjunctive combined cause for the next iteration. Our pruning algorithm detects a local-optimal solution (the global optimal solution is an NP-hard problem and cannot handle complex data efficiently).
The overall process of our proposed algorithm is illustrated in Algorithm 1.

Algorithm 1 Line 1 is the iteration for the validation of **C2** (Line 2) and **C3** (Line 5) mentioned above. Once the relation of a drug, the domain knowledge drug and the adverse effect is classified as **C2** or **C3**, this drug is no longer considered. Line 10 identifies the drug which can most significantly increase the dependence strength between d_m and s based on the conjunctive combined cause detected from previous iterations by using the greedy strategy. Line 11 removes drugs that may not be the direct cause of s. On line 13, if there is no such drug detected, then the current conjunctive combined cause is the final result. On Line 15, PMI is used to identify whether the dependence is positive or negative. For PMI validation, if $d_i \in O$ causes a lower positive dependence than the total combined causes set when not including them in O, then d_i is not considered. However, it may be considered in the future iterations, because the PMI for that drug may change with other conditions.

```
   Input: Domain knowledge drug $d_m$ and adverse effect $s$
   Other drugs $D\prime = \{d_1, d_2, ..., d_n\}$
   Threshold of dependency value $th$ = the critical value of Chi-square when significance level $\alpha = x$
   Output: Conjunctive combined cause $O$
   $O = \{\}$; //initialise output variable as an empty set
   $T = D'$; //temporary variable as a copy of $D'$
   Start
 1: for each $d_i$ do
 2:     if $chi(d_i, d_m \mid s) < th$ then
 3:         $D' = D' - \{d_i\}$;
 4:         continue;
 5:     if $chi(d_i, s \mid d_m) < th$ then
 6:         $D' = D' - \{d_i\}$;
 7:         continue;
 8: $T = \{\}$; //temporary variable for the drugs failed in PMI validating
 9: while true do
10:     $d_i = \underset{d \in \mathcal{D}' - \{O,T\}}{\operatorname{argmax}} chi(d_m, s \mid O, d)$
11:     if $chi(d_i, s \mid d_m, O) < th$ then
12:         $D' = D' - \{d_i\}$;
13:         continue;
14:     if $chi(d_m, s \mid O, d_i) <= chi(d_m, s \mid O)$ then
15:         break;
16:     if $pmi(d_m, s \mid O, d_i) <= pmi(d_m, s \mid O)$ then
17:         $T = T \cup d_i$;
18:         continue;
19:     else
20:         $O = O \cup d_i$;
21:         $T = \{\}$; //set the temporary variable to the empty set for the next loop
    End
```

3.4 Time Complexity Analysis

Next we analyse the time complexity of the proposed algorithm. The pruning step (lines 1-8) costs $O(n)$, where n is the number of drugs considered for DDI (except the drugs contained in the domain knowledge). If at least one drug is not pruned, then we need to consider the greedy step (lines 9-21). Line 10 requires n operations for the worst case in which every candidate drug will be included to form the combined causes, thus contributing $O(n)$ to the algorithm's time complexity. Lines 11 to 18 consider three constant time operations ($O(1)$). These iteration steps (lines 9-21) are considered until all drugs have been discounted (n in the worst case), or the break in line 15 occurs. Thus the greedy step costs $O(n) * [(O(n) + O(1)] = O(n^2)$. The total time complexity of the algorithm is therefore $O(n) + O(n^2) = O(n^2)$.

4 Experiments

Both synthetic and real-world data are used to empirically validate the effectiveness of the proposed method. We use synthetic data to compare our outcome with CARD – other baseline methods for causality discovery such as CBN and BCC are not considered because of the extremely high time complexity for the DDI cases considered in our experiments. Unfortunately, we cannot either compare CARD with our method on real-world data because CARD could not complete its execution due to its exponential time complexity – this highlights the major problem of CARD, which our method aims to resolve. The hardware used for the experiment was a server with an Intel i7-6700 CPU and 16G of RAM.

4.1 Domain Knowledge

Our method relies on the availability of domain knowledge about drugs that are known to cause the target adverse effect. We acquire this data from DailyMed, a database of trustworthy adverse effects extracted from official drug labels[5].

4.2 Real-World Data

We use approximately 300,000 patient records collected from FAERS. The extracted data contains missing values and duplicates, as noted previously by others [14, 3], and thus we preprocess the data according to previous works. Duplicate reports are removed when they contain at least eight drugs or adverse effects, and all drugs, adverse effects and patient demographic information are the same as in [3]. Reports with missing adverse effect are not considered. Only drugs and adverse effects that occur in at least 5 reports are included. Two FAERS attributes are considered in our experiments: 'adverse event' and 'drug name', having approximately 10,000 and 40,000 values, respectively. For drug names, we conflate different names of the same drug to a unique identifier using

[5] https://dailymed.nlm.nih.gov/dailymed/

a method proposed by Banda et al. [2]. We consider all types of adverse effects to provide all available ranges of adverse symptom severity.

In this dataset, a patient may have multiple records indicating that they may take many medicines to cure one or more diseases, and report several adverse effects: it is unclear what drug or combination of drugs has provoked which adverse effect(s). That is, the true causal DDICAE cannot be directly identified from FAERS data. In fact, for each report, the recorded set of drugs usage and adverse effects might be consistent with many possible cases: e.g., (1) some DDIs caused many adverse effects, (2) there are more than one DDICAE, (3) some drugs are used to cure adverse effects that are not the DDI, (4) some drugs do not cause any reported adverse effects, etc. Because of this, the reliance on methods that only depend on correlation between datapoints/features may identify relations between drugs and adverse effects that are actually not causal DDICAE effects – and thus fail to reveal correct insight relationships.

To evaluate on real-world data, the top ten causal drug-drug interactions results discovered using DCD are selected, based on the ranking from the dependence strength of the relations and their positive dependence measurements. To evaluate the prediction correctness, two reliable pharmaceutical drug-drug interaction databases, MedicinesComplete (MedComp)[6] and Drugs.com[7], are selected. The prediction results that do not match with the ground truth in the databases are not necessarily incorrect: the causality might not be confirmed or may have not been yet discovered by clinical or biological methods yet. In our evaluation, the label 'Not Found' is used for a drug-drug interaction that is not found in these databases.

4.3 Synthetic Data

The DDICAE predictions from real-world FAERS data only offer limited ground truth data and thus do not allow for a complete, reliable evaluation. Thus, in addition to FAERS data, we generate synthetic evaluation data using Tetrad[8], a tool widely used in previous studies to generate causal Bayesian graphs for evaluation [1, 11]. Tetrad generates the directed acyclic causal Bayesian graph with known causal paths between variables. We generate three groups of graphs, the graphs with 50, 100 and 200 variables, to show the applicability of our method. In our setup, each group contains 10 random graphs and each variable is randomly assigned its directed causal links to connect to other nodes (between 0 to 7 causal links). Each graph contains 10,000 records, no loops, and any two nodes can only have one edge between each other. The conditional probability tables of the causal Bayesian networks are also randomly generated. Binary data

[6] MedicinesComplete published in Pharmaceutical Press and the Royal Pharmaceutical Society: https://www.medicinescomplete.com/mc/alerts/current/drug-interactions.htm

[7] Data sources from Micromedex, Multum and Wolters Kluwer databases: https://www.drugs.com/drug_interactions.php

[8] http://www.phil.cmu.edu/tetrad/

Table 1. The average Chi-square(%) and average PMI (log scale) of predicted DDI and the adverse effect compared to d_m and s, by DCD for variable sizes 50,100 and 200 and FARES.

Method	50	100	200	FAERS
Chi-square	1,349.77%	1,167.22%	947.09%	172.22%
PMI (log)	0.13	0.14	0.12	0.42

for variables in all records are generated based on the conditional probability tables using the Bayesian Instantiated Model. The generated data is preprocessed in the same manner as FAERS data. In the case of domain knowledge from synthetic data, for each child node having at least one parent, one parent is randomly selected as domain knowledge.

To illustrate the effectiveness of our method on the synthetic data, we measure the precision of discovering drug-drug interactions. In our case, precision is defined as $TP/(TP + FP)$, where TP is the number of correct predicted causal links, and FP is the number of incorrectly predicted causal links.

Recall cannot be used for evaluation in both synthetic and real-world data because we do not know which predicted groups of drugs are the true conjunctive combined causes of the adverse events. In place of recall, we use Chi-squared value and PMI. The drug-drug interaction with the highest Chi-squared value and positive PMI is considered significant (with respect to d_m and s). We use this method for a recall-oriented evaluation for both synthetic and real-world data.

5 Results and Discussion

First, we evaluate whether DCD produces a higher quality conjunctive-combined causes compared to inputted domain knowledge and target effects from synthetic data and FAERS. In Table 1, the results of DCD are compared using domain knowledge and the target adverse effects. The evaluation shows that the outcomes have stronger dependence relationship in both synthetic data and causal drug-drug interaction and its adverse effect in FARES measured by the Chi-square. The strength of positive dependence relationships measured with the PMI in our outcomes is also stronger. These results imply that the interaction outcomes causing related effects from both synthetic data and real-world data have a higher probability to be a true cause outcome compared to the domain knowledge and its adverse effect.

Table 2 presents the results of drug-drug interactions and adverse effects from FARES as predicted by DCD. The top ten predictions with the highest positive dependence are selected to be validated with the two pharmaceutical databases, MedComp and Drug.com. Note that DCD can detect DDI cases that have more than 2 drugs (eg. such as LIPITOR + CRESTOR + NEXIUM → *Hypertension*). This table shows that nine out of ten drug-drug interactions

Table 2. Comparison between drug-drug interactions and their adverse effect predicted by DCD and found in selected reliable pharmaceutical databases (D→Drugs.com and M→MedComp).

DDI predicted by DCD	Adverse Effect	Found in Database
ENBREL + HUMIRA	Sepsis	M, D
PREDNISONE + ENBREL	Infections	D
ENBREL + REMICADE	Drug ineffective	D
ENBREL + HUMIRA	Pain	M, D
METHOTREXATE + ORENCIA	**Drug ineffective**	**Not Found**
ENBREL + PLAQUENIL	Drug ineffective	M, D
TEMAZEPAM + GABAPENTIN	Dizziness	D
PREDNISONE + ENBREL	Pain	D
LIPITOR + CRESTOR + NEXIUM	Hypertension	D
ENBREL + METHOTREXATE	Fatigue	D

Table 3. The average precision (%) on synthetic data of DCD compared to CARD for variable sizes 50,100 and 200. Highest values are indicated in **bold**.

Method	50	100	200
DCD	**91.67**	**86.96**	**83.61**
CARD	84.61	76.74	59.01

Table 4. The average computation time (seconds) of DCD compared CARD for variable sizes 50, 100 and 200. Lowest values are indicated in **bold**.

Method	50	100	200
DCD	**0.36**	**2.95**	**10.95**
CARD	55.84	314.17	1,918.78

and their adverse effect results from FARES predicted by DCD are found in Drugs.com, and three of them are found from both of databases. Only one case, METHOTREXATE + ORENCIA → *Drug ineffective*, is not found in both databases. This outcome can be a candidate for biological tests to validate the causality in the future.

Next, we evaluate the pruning on synthetic data. DCD successfully prunes 96.99% of unrelated interactions, while CARD only removes 17.31%. In addition, the remaining interactions using DCD are more likely to have higher probability to be the valid conjunctive combined cause that causes adverse effect or DDI-CAE. The results of these experiments are present in Table 3. Here, the more variables in the synthetic data indicates a higher complexity of the causal graph (i.e., closer to the real-world data). The results of DCD outperform those of CARD in precision using the synthetic data. The computation time of DCD is also lower than that of CARD, which is described in Table 4. When compared to CARD, our method tends to have higher probability to predict causality with higher precision and lower computation time.

To compare the dependence strength of our approach with CARD, Chi-squared value is used to measure dependence strength, and PMI is used to measure the positive dependence relationship, as shown in Table 5. When using synthetic data, we outperform CARD in all measures tested. When using real-

Table 5. Average Chi-squared value and PMI of predicted DDI with adverse effect of DCD compared to CARD for variable sizes 50,100 and 200 and FARES.

	Average Chi-squared value				Average PMI			
Method	50	100	200	FARES	50	100	200	FARES
DCD	1,478.84	3,255,886.03	7,192.55	833.34	0.32	0.30	0.20	1.25
CARD	36.31	4,312.94	569.13	-	(-0.21)	(-0.81)	0.02	-

world data, DCD also performs effectively; however, the number of variables are too large for CARD to execute because of its exponential time complexity. The results of CARD show the negative dependence (negative value in the bracket) because it uses mutual information to measure the dependence (i.e., it ignores positive and negative dependence). Therefore, CARD tends to be not as effective at discovering causal DDICAE because it cannot identify co-occurrence DDICAE (i.e., it relies on the positive dependence relationship of related drugs).

DCD is the only CBN and BCC method used to discover DDICAE and is state-of-the-art in terms of both effectiveness and efficiency (when compared to CARD). DCD does not only find causality in DDICAE but also confirms the causal direction using domain knowledge and our proposed pruning steps. As highlighted by our evaluation on real-world data, DCD is also more likely to discover true causal DDICAE and to be an effective and efficient method to support decisions for pharmaceutical and medical experts.

6 Conclusion

Causality discovery in drug-drug interaction and adverse effect is an important task for health care decision support. Traditional methods to confirm causality from drug-drug interaction and adverse effects, such as biological experiments, are difficult, complicated and expensive. Computational methods, such as CBN (e.g., PC-Algorithm), BCC methods (e.g., Markov Blanket), are effective to discovery causality; however, they generally suffer from exponential time complexity, and BCC is affected by the causality direction ambiguity problem. CARD is the only CBN method to effectively discover causality in drug-drug interaction and adverse effect. However, it does not perform well with real-world data and also suffers from its exponential time complexity nature.

In this paper, we have proposed the Domain-knowledge-centred Causality Discovery algorithm (DCD) that can discover causality from drug-drug interactions and adverser effect. Advantages of our algorithm include:

- Domain knowledge is used effectively as a guide to discover causality, and it can confirm the causal direction when used with our proposed pruning steps.
- Computation time is reduced by pruning most of the irrelevant drugs of the target DDICAE (Drug-Drug interaction Causing Adverse Effect) by using the proposed evidence-based DDICAE structure (relating to the domain knowledge) integrated with conditional independence.

- The discovered DDICAEs are meaningful because they are the co-occurrence DDICAEs represented by the positive dependence values (unlike e.g., CARD, that does not exploit such co-occurrence properties).
- The discovered DDICAEs include those that can and cannot be detected using the V-structure property, unlike current state-of-the-art methods such as CARD, that solely rely on the V-structure, whose underlying data assumptions are in fact not necessarily satisfied in real-world data.
- The time complexity of the proposed method is polynomial ($O(n^2)$) because our algorithm applies a greedy algorithm to find causality, and this is a speed-up compared to current state-of-the-art approaches (e.g., CARD).

However, our method provides a locally optimal solution and may not find all drugs in the drug-drug interaction causing the adverse effect. In addition, our method cannot be used when no domain knowledge exists. However, the outcomes are still helpful to reduce the cost of drug-drug interaction discovery and as a decision support for doctors when they prescribe medications to patients.

Acknowledgements The authors would like to acknowledge Meng Wang from Southeast University, China, and Mingyang Zhong from The University of Queensland, Australia, for their input on the initial stages of this work.

References

1. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. Journal of Machine Learning Research **11**(Jan), 171–234 (2010)
2. Banda, J.M., Evans, L., Vanguri, R.S., Tatonetti, N.P., Ryan, P.B., Shah, N.H.: A curated and standardized adverse drug event resource to accelerate drug safety research. Scientific data **3**, 160026 (2016)
3. Cai, R., Liu, M., Hu, Y., Melton, B.L., Matheny, M.E., Xu, H., Duan, L., Waitman, L.R.: Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artificial intelligence in medicine **76**, 7–15 (2017)
4. Chickering, D.M., Heckerman, D., Meek, C.: Large-sample learning of bayesian networks is NP-hard. Journal of Machine Learning Research **5**(Oct), 1287–1330 (2004)
5. Hansen, M.L., Sørensen, R., Clausen, M.T., Fog-Petersen, M.L., Raunsø, J., Gadsbøll, N., Gislason, G.H., Folke, F., Andersen, S.S., Schramm, T.K., et al.: Risk of bleeding with single, dual, or triple therapy with warfarin, aspirin, and clopidogrel in patients with atrial fibrillation. Archives of internal medicine **170**(16), 1433–1441 (2010)
6. Harpaz, R., Perez, H., Chase, H.S., Rabadan, R., Hripcsak, G., Friedman, C.: Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clinical Pharmacology & Therapeutics **89**(2), 243–250 (2011)
7. He, L., Yang, Z., Lin, H., Li, Y.: Drug name recognition in biomedical texts: a machine-learning-based method. Drug discovery today **19**(5), 610–617 (2014)

8. Jin, Z., Li, J., Liu, L., Le, T.D., Sun, B., Wang, R.: Discovery of causal rules using partial association. In: IEEE 12th International Conference on Data Mining. pp. 309–318. IEEE (2012)
9. Lazarou, J., Pomeranz, B.H., Corey, P.N.: Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama **279**(15), 1200–1205 (1998)
10. Li, J., Le, T.D., Liu, L., Liu, J., Jin, Z., Sun, B.: Mining causal association rules. In: IEEE 13th International Conference on Data Mining (Workshops). pp. 114–123. IEEE (2013)
11. Li, J., Ma, S., Le, T., Liu, L., Liu, J.: Causal decision trees. IEEE Transactions on Knowledge and Data Engineering **29**(2), 257–271 (2017)
12. McHugh, M.L.: The chi-square test of independence. Biochemia medica: Biochemia medica **23**(2), 143–149 (2013)
13. Neapolitan, R.E., et al.: Learning bayesian networks, vol. 38. Pearson Prentice Hall Upper Saddle River, NJ (2004)
14. Norén, G.N., Orre, R., Bate, A., Edwards, I.R.: Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery **14**(3), 305–328 (2007)
15. Pearl, J.: Causality. Cambridge University Press (2009)
16. Qin, X., Kakar, T., Wunnava, S., Rundensteiner, E.A., Cao, L.: MARAS: Signaling multi-drug adverse reactions. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1615–1623. ACM (2017)
17. Quinn, D., Day, R.: Drug interactions of clinical importance. Drug safety **12**(6), 393–452 (1995)
18. Rodrigues, David, A.: Drug-drug interactions. CRC Press (2008)
19. Van Puijenbroek, E.P., Egberts, A.C., Meyboom, R.H., Leufkens, H.G.: Signalling possible drug–drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. British journal of clinical pharmacology **47**(6), 689–693 (1999)
20. Waldmann, M.R., Martignon, L.: A bayesian network model of causal learning. In: Proceedings of the 20th annual conference of the Cognitive Science Society. pp. 1102–1107 (1998)

Appendix I: Terminology used in this Paper

BCC: *Bayesian Constraint-based Causality*
CARD: *Causal Association Rule Discovery*
CBN: *Causal Bayesian Network*
DCD: *Domain-knowledge-driven Causality Discovery*
DDICAE: *Drug-Drug Interactions Causing Adverse Effects*
DDI: *Drug-drug interactions*
FAERS: *Food and Drug Administration (FDA) adverse event report system*