Computers in Biology and Medicine 155 (2023) 106636

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# CauRuler: Causal irredundant association rule miner for complex patient trajectory modelling

Guillem Hernández Guillamet [a,b,c,*], Francesc López Seguí [b,c], Josep Vidal-Alaball [d,e,f], Beatriz López [a]

[a] eXiT Research Group, Universitat de Girona (UdG), EPS - Edifici P-IV, Carrer Universitat de Girona, 6, Girona, 17003, Catalunya, Spain
[b] Assistance strategy management. Hospital Germans Trias i Pujol, (ICS), Carretera de Canyet, Badalona, 08916, Catalunya, Spain
[c] Research Group on Innovation, Health Economics and Digital Transformation, Institut Germans Trias i Pujol (IGTP), Cami de les Escoles, Badalona, 08916, Catalunya, Spain
[d] Health Promotion in Rural Areas Research Group. Gerencia Territorial de la Catalunya Central, ICS, Carrer Pica d'Estats, 13-15, 08272, Sant Fruitos de Bages, Catalunya, Spain
[e] Unitat de Suport a la Recerca de la Catalunya Central, Fundacio Institut Universitari per a la Recerca a l'Atencio Primaria de Salut Jordi Gol i Gurina, Gran Via de les Corts Catalanes, 587, 08007, Barcelona, Catalunya, Spain
[f] Faculty of Medicine, University of Vic-Central University of Catalonia, Ctra. de Roda, 70, 08500, Vic, Catalunya, Spain

## ARTICLE INFO

*Keywords:*
Causal inference
Association rules
Non-redundant associations
Health data mining

## ABSTRACT

**Background and Objectives.** Discovering causal associations between variables is one of the main goals of clinical trials, with the ultimate aim of identifying the causes of specific health status. Prior knowledge of causal paths could help ensure patients do not develop the resultant conditions. In recent years, thanks to the enormous amount of health data stored with the support of digital tools, attempts have been made to employ Machine Learning to infer causality. Those methodologies suffer from some deficiencies in controlling cofounders when analysing causality, as well as providing causal rules general enough to be useful in healthcare practice. Conversely, this work presents and evaluates CauRuler, a new approach to deal with causality from association rules. The proposed approach uses a pruning strategy to reduce the association rule set, which does not compromise the causality learning capability of the algorithm. This behaviour makes the algorithm suitable for exploiting large health databases with thousands of patients and medical instances. CauRuler can control a larger number of confounders than other proposals, bringing robustness to causal analysis and avoiding the identification of spurious associations. Additionally, the method generalizes causality using anti-monotone properties to obtain complex and general causal paths. The method can target correct causal associations in complex medical databases with retrospective data.

**Method:** CauRuler extends association rule mining with an irredundancy property so that the set of rules learnt is reduced in size and generalized. General association rules, conformed by fewer items, enable controlling more confounding variables to verify, with more statistical evidence on available data, if they represent causal paths in patient disease trajectories.

**Results:** CauRuler has been tested on a complex real medical database (3,5 M visits to the primary care services between 2019 and 2020, and controlling over 15.000 different variables including diagnoses and demographic and other clinical patient data). The reduction of the rule set achieved by the pruning strategy goes from 7.732 to 2.240 rules, from which 46 have been found to have causality relationships in the patient trajectories, and generalized to 14 rules tested as true causal relationships thanks to the confounding analysis. These rules have been validated by clinicians with the support of a graphical map. The obtained causal paths control in average of 906 confounder variables, retrieving robust results.

**Conclusions:** Causal relationships enable predicting causal paths between health conditions according to patient trajectories. Knowing these causal paths is crucial for understanding and preventing the appearance or worsening of diseases in patients. CauRuler, with high demanding thresholds, has proven its efficiency and effectiveness in targeting previously known causal associations between diagnoses, reaching consensus in the medical community. Softening these thresholds should help target interesting general causal paths.

* Correspondence to: Hospital Germans Trias i Pujol, Carretera de Canyet, Badalona, 08916, Spain.
*E-mail address:* ghernandezgu.germanstrias@gencat.cat (G.H. Guillamet).

https://doi.org/10.1016/j.compbiomed.2023.106636
Received 12 December 2022; Received in revised form 25 January 2023; Accepted 4 February 2023
Available online 9 February 2023

## 1. Introduction

Some of the major milestones in modern medicine were achieved immediately following the identification of causal associations between clinical conditions, agents and treatments. The identification of contaminated water as the causal source of cholera in the late nineteenth century, for example, represented a breakthrough and went on to save many lives. Nevertheless, when dealing with causality, effects are typically determined by a series of causes interacting to a greater or lesser degree; not being exclusive between them. A classic example is the failure to identify tobacco consumption as one of the major causes of lung cancer until as late as 1964 [1]. Non-smokers can experience lung cancer and regular smokers may never suffer from it. This can be explained by the fact that smoking is not the sole cause of lung cancer, and it does not represent 100% of those who suffer from it. Other causal factors such as genetic predisposition and pollution also play a crucial role. Hence, we can find necessary causes (water contaminated by faecal matter → Cholera) as well as sufficient causes (smoking → lung cancer) [2]. In other words, when the same causes are present, the same effects are produced; even so, this does not imply that an effect is always produced and that there is a causative relationship.

Historically, Randomized Controlled Trials (RCTs) have been the most commonly used research method for identifying causality in the health field. They are based on detecting as much experimental data supporting the statistical dependency between $X$ and $Y$ to conclude the first causally influences the latter $(X \rightarrow Y)$ [3]. However, this approach usually has a series of limitations. RCTs might not be logistically feasible (need of a minimum number of subjects in trials [4]), ethical (if causes are hazardous for the subjects), financially worthwhile or even theoretically possible.

The definition of causality gave rise to many philosophical and mathematical problems for some time. Pearl and Verma initially proposed the basis of the semantics in the mathematical field to define causality [5]. Employing these semantics, graphical causal modelling algorithms were designed. The majority of these methods are based on Bayesian networks and other probabilistic causal modelling approaches. Both Bayesian networks and causal networks represent a breakthrough in the field of causality, having reached a consensus on their effectiveness. Nevertheless, Bayesian learning is a computationally expensive technique, underperforming when scaling to large data volumes, as the ones in the health field [6]. While mathematically robust, they are of no use when it comes to extrapolating them to real scenarios [7]. Other methods aim to overcome the aforementioned computational issues by learning partial causal networks [8,9].

Machine Learning (ML) and Artificial Intelligence (AI) have enabled the development of alternative methods based on the analysis of retrospective data stored in healthcare systems. In particular, Association Rule miners (ARMs) identify associations and correlations between variables in a transactional database structured as association rules with the form "if-then" $(X \rightarrow Y)$ [10]. Nevertheless, correlation implies association, but not necessarily causation [11]. Hence, we cannot assume the association rules targeted by those methods have a causal nature. However, all causal associations by definition are associations [12]. With these prerequisites, some proposals of algorithms have attempted to detect those causal structures from association rules [13,14]. Each resulting association is handled as an independent observational respective study to detect if there is sufficient statistical evidence to support causality while controlling variables which can play the role of confounders (as age, and gender, typically are). This methodology presents an improvement concerning RCTs since it uses retrospective observational data rather than prospective data, and unlike most of the bayesian approaches, can be escalated to large data volumes. Nevertheless, those methods might lose some of the causal associations thus observing partial causal networks.

Although being interesting approaches to target causality, ARMs tend to suffer from high computational costs, with significant research carried out to overcome the issue [15,16]. Moreover, there exists a limitation related to the *output parametrization tradeoff*: algorithms depend on different parameters whose settings condition the size of the rule set obtained. If they are over-constrained, the size of the rule set is too small, discovering too evident associations. Conversely, softening the thresholds yields a basis of rules which is too large to be interpreted. This calls for a certain degree of expertise and sheer luck to set the parameters correctly to find the rules governing the previously unknown data. Some methodologies tried to improve the aforementioned aspect by introducing a *novelty notion*. The objective is performing a pruning of the resulting rule set through the use of other parameters which evaluate the *novelty* or *redundancy* of the association rules to achieve a minimum sized non-redundant set of rules [17–19]. Thanks to this methodology, a minimum set of rules is obtained which can be easily interpretable without the loss of too much information.

This research is concerned to develop an algorithm able to automatically discover causal rules while trying to avoid information loss in medical databases. The algorithm suggested in this paper, CauRuler, is based on the procedure from [13]. The main contributions of the algorithm are the use of a pruning method to reduce the association rule set size, which does not compromise the causality learning capability of the algorithm. This behaviour makes the algorithm suitable for exploiting larger health databases. Moreover, CauRuler can control a larger number of confounders than other proposals, bringing robustness to causal associations and avoiding the identification of spurious associations. Additionally, the method generalizes causality using anti-monotone properties to obtain complex and general causal relations. Finally, the method outputs causal maps where interactions between different generalized causal patient trajectories are shown thus facilitating the clinician's interpretation of the sufficient causes for certain clinical conditions.

Experimentation is carried out on a complex database from a real clinical environment involving the diagnoses of 400.000 patients during the years 2019 and 2020 (including part of the covid-19 period). The resulting causal paths were evaluated by different physicians to ensure their suitability and the effectiveness of the model to target known or feasible causal associations for high demanding thresholds.

### 1.1. Related work

In the field of health data mining, causal structures are usually captured using bayesian approaches. There are multiple examples in the literature of Bayesian networks being used to develop causal maps in medical databases [20]. Others are specific to a particular condition [21]. Causal association paths might not be only interesting at the patient level to prevent diseases, but also may be used to design complex models to improve understanding in patient evolution [22].

ARMs have been used to identify patterns for specific diseases [23], comorbidities related to a target disease [24] or multi-morbidity patterns [25]. Others aimed to find patterns in large databases of Electronic Health Records (EHR) involving a collection of variables such as diseases, symptoms and drugs [26]. However, those methods output associations expressed through rules of the form "if-then", that do not necessarily represent causation. Alternatively, other research complements ARMs with algorithms that enable the identification of the association structures that are causal to understand patient trajectory in health [13,14].

## 2. Methods and materials

Fig. 1 illustrates the CauRuler workflow proposed in this paper to mine causal associations. The algorithm can be divided into three parts: data preprocessing, associations mining and the causality study of the former. Appendix A at the end of the paper contains the variables and parameter definitions which can be helpful for the reader to understand the algorithms and mathematical definitions in the following sections.

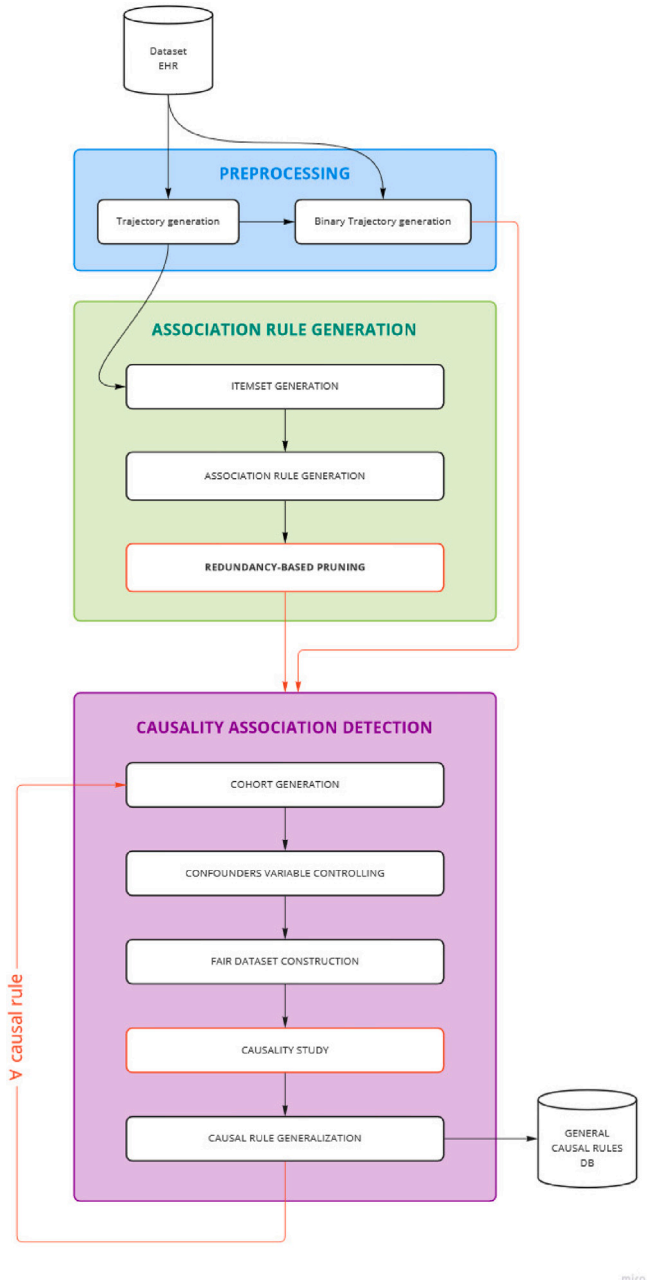


**Fig. 1.** CauRuler methodology workflow. Trajectories and the binary representations of the former are constructed from the same Electronic Health Records (EHR) database. Raw trajectory is used for rule mining while binary representation is used to test the causality nature of each association.

## 2.1. Preprocessing

The raw data used in the method are EHR databases with health information representing patients' medical histories. EHRs consist of real-time, patient-centred records and contain diagnoses, medications, treatment plans, allergies, etc. The data needs to be transformed before ML algorithms can be applied. This method performs two preprocessing: generation of clinical trajectories (from which we obtain a representation of the EHR in the format of medical trajectories to apply rule mining) and the binary representation of the trajectories (which allow the causality study and the controlling of confounders).

### 2.1.1. Trajectory generation

Rule miners act on transactional databases to mine associations between products. In a medical context, we identify transactions as *patient clinical trajectories* $T = \{t_{p_1}, \dots, t_{p_M}\}$ given a set of patients $P = \{p_1, \dots, p_M\}$. In the aforementioned frameworks, the products represent clinical instances such as procedures, drugs, results and diagnoses. In the matter at hand, clinical trajectories are completely formed by diagnoses defined in a tuple $D = \langle d_1, \dots, d_N \rangle$, where $N$ represents the total number of distinct diagnoses. For example, $D$ could represent the standard ICD-10 ontology [27].

**Definition 1** (*Clinical Trajectory*)**.** A clinical trajectory $t$ of patient $p_i$ is formed by all the diagnoses $d_i \in D$ coded for patient $p_i$ in a determined period of time for a EHR database. Trajectories may have different lengths and contain diagnose repetitions.

$$t_{p_i} = \langle d_1^{p_i}, d_2^{p_i}, \dots, d_{n_{p_i}}^{p_i} \rangle \text{ on } d_j^{p_i} \in D$$

### 2.1.2. Binary trajectory generation

Given a set of patients $P$ and their EHR records in a database, and a set of trajectories $T$, this pre-processing step generates binary trajectories of patients $T^b = \{t_{p_1}^b, \dots, t_{p_M}^b\}$ which contain all the information required for a causal analysis of patients' trajectories.

The patient binary clinical trajectory is a binary representation of the clinical trajectory of the patient combined with other *clinic interest variables* ($V = \langle v_1, \dots, v_K \rangle$) that could be relevant for her health, such as sex, age and level of medical coverage. $V$ variables play an important role since they can act as confounders (it is crucial that they are equally distributed between the control and exposed cohorts, healthy and non-healthy cohort respectively).

To facilitate the posterior controlling of confounder variables the clinical data needs to be preprocessed to reach a binary representation $V^b$ of the $V$ variables. If $v_i \in V$ is a numeric variable, it is categorized into a new discrete variable $v_i^c$ taking values from a predefined set. Hereafter, all categorical variables $v_i \in V$ i $v_i^c$ are binarized resulting in a set of variables $V^b = [v_1^b, \dots, v_{K^b}^b]$, with $v_j^b \in \{0, 1\}$. This is commonly known as one-hot-encoding [28,29]. For example, the variable age is discretized in 10 year intervals generating 10 binary variables, one per discrete value. Each patient will only present 1 in one of this set of discretized variables.

**Definition 2** (*Binary Clinical Trajectory*)**.** Given a trajectory $t_{p_i}$ for patient $p_i$, and their binary clinical values, $v_1^{(b)p_i}, \dots, v_{K^b}^{(b)p_i}$, the binary clinical trajectory $t^b$ of patient $p_i$ has the form $t_{p_i}^b = \langle b_1^{p_i}, \dots, b_N^{p_i}, b_{N+1}^{p_i}, \dots, b_{N+K^b}^{p_i} \rangle$ such that:

- $b_j^{p_i} \in \{0, 1\}$
- $\forall j \in \{1, \dots, N\}, b_j^{p_i} = \begin{cases} 1 & if \exists d_l^{p_i} \in t^{p_i}, pos(d_l^{p_i}, D) = j \\ 0 & otherwise \end{cases}$.
- $\forall j \in \{N+1, \dots, N+K^b\}, b_j^{p_i} = v_j^{(b)p_i}$

Where $pos(d_l^{p_i}, D)$ is a function that returns the position of the diagnosis $d_l^{p_i}$ in $D$.

Binary trajectories present a stable length predefined by the length of the dimension of clinical diagnoses in an EHR database ($N$) and the set of variables of clinical interest $K^b$. The variables set used to construct the binary trajectory dataset is defined by $B = D \cup V^b$.

## 2.2. Association rule mining

Given a set of clinical trajectories $T = \{t_{p_1}, \dots, t_{p_M}\}$, this mining step obtains association rules of the form: $r : X \rightarrow Y$, where $X$ and $Y$ are subsets of diagnoses in $D$, meaning those trajectories presenting diagnoses in $X$ tend to present those in $Y$, therefore being *frequent patterns*. This process is achieved by a three-stage process. First, frequent itemsets are generated; second, the relationship between the elements

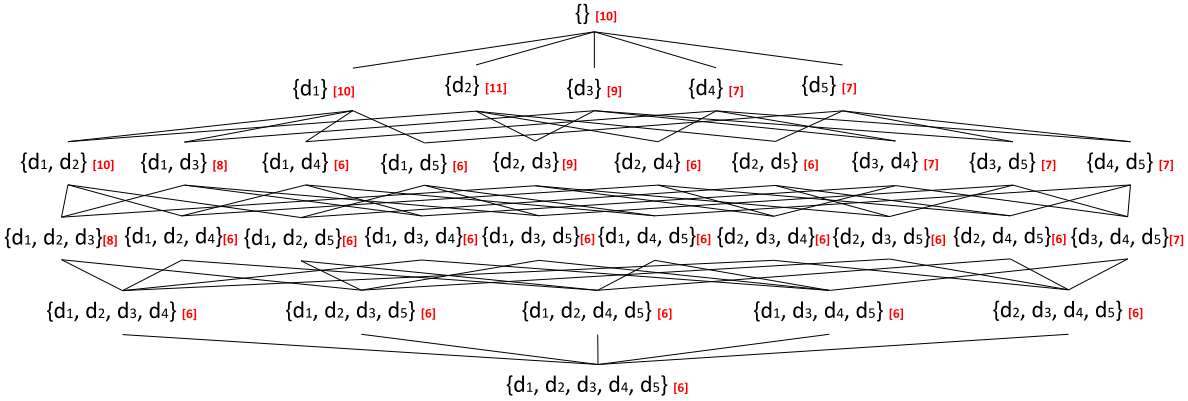


**Fig. 2.** Lattice of possible itemsets for a set of $N = 5$ diagnosis of $D$. Between brackets, number of itemset appearances in the example in Table 1.

**Table 1**
Trajectory definition of patients in the Example 1.

| Trajectory | Patients |
|---|---|
| $\langle d_1, d_2, d_3, d_4, d_5 \rangle$ | $p_1, p_2, p_3, p_4, p_5, p_6$ |
| $\langle d_1, d_2, d_3 \rangle$ | $p_7, p_8$ |
| $\langle d_1, d_2 \rangle$ | $p_9, p_{10}$ |
| $\langle d_3, d_4, d_5 \rangle$ | $p_{11}$ |
| $\langle d_2, d_3 \rangle$ | $p_{12}$ |

of the frequent itemsets is shaped by means of association rules. Finally a pruning strategy is applied to reduce the quantity of generated rules. As a result, a set of rules $R = \{r_1, \ldots, r_n\}$ is obtained.

#### 2.2.1. Itemset generation

This step consists in finding frequent diagnoses sets in $T$: meaning they are at least present $\tau$ times in $T$, or what is the same, have a minimum *support* $\tau$.

**Definition 3** (*Diagnosis Set Support*)**.** The support $s(Z,T)$ of a set of diagnoses $Z \subseteq D$ regarding a set of patient trajectories $T$ is the cardinality of the set of clinical trajectories in $T$ containing $Z$: $s(Z,T) = |\{t_{p_i} \in T | Z \subseteq t_{p_i}\}|$.

**Definition 4** (*Frequent Set Of Diagnoses*)**.** A frequent set of diagnoses or frequent itemset $Z \subseteq D$ with regard to a database $T$ is frequent if $s(Z,T) \geq \tau$, where $\tau$ is a given support threshold.

The procedure used to generate all frequent itemsets follows the *apriori* basic approach [10]. First, all one-length itemsets are generated, which means frequent itemsets containing a single diagnosis among all the possible diagnoses in $D$. Frequent one-length itemsets (based on $\tau$) are retained. Second, two-length itemsets are generated, by expanding the one-length frequent itemsets with a diagnosis in $D$. Again, only frequent two-length itemsets are retained. And so on until reaching the $N$ length itemsets or no frequent itemsets are found.

**Example 1.** Consider an example that will be used in the upcoming sections. The universe of diagnoses $D$ includes the five diagnoses $d_1, d_2, d_3, d_4, d_5$ ($N = 5$). The dataset consists of 12 patients: six of which include all elements in $D$; two more consists of $\langle d_1, d_2, d_3 \rangle$, again two trajectories consist of $\langle d_1, d_2 \rangle$, and then one trajectory consists of $\langle d_3, d_4, d_5 \rangle$ and another one consists of $\langle d_2, d_3 \rangle$. Table 1 define the trajectories.

All the possible combinations of items are $2^N - 1$ (see an example represented by a lattice in Fig. 2). The requirement of a given support $\tau$ reduces the number of possible itemsets which can be generated according to the available trajectories.

In Fig. 2, for a $\tau = 0.6$, 9 frequent itemsets are obtained:

$$\{d_1\}, \{d_2\}, \{d_3\}, \{d_4\}, \{d_5\}, \{d_1, d_2\}, \{d_1, d_3\}, \{d_2, d_3\}, \{d_1, d_2, d_3\}$$

The later ones are the more interesting as representing combinations of different diagnoses.

#### 2.2.2. Association rule generation

In this step, the relation among the elements of a frequent itemset $Z$ is shaped through association rules $r : X \rightarrow Y$, where $Z = X \cup Y$. Based on the matter at hand, in the following sections association rules will be called *diagnosis association rules*.

**Definition 5** (*Diagnosis Association Rule*)**.** $r : X \rightarrow Y$ is defined from a frequent itemset $Z = X \cup Y$, given a dataset of patient trajectories $T$, if confidence $c(r,T) \geq \gamma$, being $c(r,T) = \frac{s(X \cup Y, T)}{s(X,T)}$, and $\gamma$ the confidence threshold.

The confidence $c(r,T)$ of a rule in a dataset of patient trajectories $T$ is an empirical approximation to the conditional probability (how frequent is $Y$ among all transactions containing $X$). Observe that the two rules $X \rightarrow Y$ and $Y \rightarrow X$ can be derived from the same itemset $Z = X \cup Y$. Nevertheless, it is expected that the confidence threshold helps to reduce the amount of rules to be obtained by a rule miner.

**Example 2** (*Continuation Example 1*)**.**
Table 2 shows the different rules obtained from the dataset defined in Example 1, with $\tau = 0.6$ and $\gamma = 0.9$. Only 5 rules reach the confidence and support thresholds in our dataset.[1]

#### 2.2.3. Redundancy-based pruning

This step aims to reduce the number of associations, ideally to those more prone to be causal. Through the causality analysis of a reduced set of rules, which is the more computationally expensive part, the efficiency of the algorithm improves and allows the analysis of more complex and bigger databases. Other propositions such as Bayesian networks use a brute-force approach to discover all causal associations, though are computationally much more expensive [30–32]. CauRuler heuristic is to keep the most general rules to discover causality. Using this heuristic the algorithm can test a smaller set of associations while still detecting a vast amount of causal ones. It loses some specificity – probably not discovering the whole causal paths – in return to designing a method capable of analyzing highly complex databases.

To perform this selection, the CauRuler algorithm uses the confidence boost $cb$ proposed by Balcazar [19]. This parameter evaluates

[1] Rules of the form $\emptyset \rightarrow \{d_1, d_2\}$ and itemsets of single diagnoses are not considered since do not bring enough knowledge.

**Table 2**
Confidence evaluation of rules targeted in Example 1 for $\tau = 0.6$.

| Association | Support | Rules | Confidence |
|---|---|---|---|
| $\{d_1, d_2\}$ | $10/12 = 0.83$ | $d_1 \rightarrow d_2$ | $10/10 = 1.0$[a] |
| | | $d_2 \rightarrow d_1$ | $10/11 = 0.91$[a] |
| $\{d_1, d_3\}$ | $8/12 = 0.67$ | $d_1 \rightarrow d_3$ | $8/8 = 1.0$[a] |
| | | $d_3 \rightarrow d_1$ | $8/9 = 0.88$ |
| $\{d_2, d_3\}$ | $9/12 = 0.75$ | $d_2 \rightarrow d_3$ | $9/11 = 0.82$ |
| | | $d_3 \rightarrow d_2$ | $9/9 = 1.0$[a] |
| $\{d_1, d_2, d_3\}$ | $8/12 = 0.66$ | $d_1 \rightarrow d_2, d_3$ | $8/10 = 0.8$ |
| | | $d_2 \rightarrow d_1, d_3$ | $8/11 = 0.7$ |
| | | $d_3 \rightarrow d_1, d_2$ | $8/9 = 0.88$ |
| | | $d_1, d_2 \rightarrow d_3$ | $8/10 = 0.8$ |
| | | $d_1, d_3 \rightarrow d_2$ | $8/8 = 1.0$[a] |
| | | $d_2, d_3 \rightarrow d_1$ | $8/9 = 0.88$ |

[a]Rules reaching confidence threshold $\gamma = 0.9$.

**Table 3**
Rules $X' \rightarrow Y'$ matching condition $X' \subseteq X$ and $Y \subseteq Y'$ in Example 1.

| Rule | Confidence |
|---|---|
| $d_1 \rightarrow d_2, d_3$ | **0.80** |
| $d_1 \rightarrow d_2, d_3, d_4$ | 0.60 |
| $d_1 \rightarrow d_2, d_3, d_5$ | 0.60 |
| $d_1 \rightarrow d_2, d_3, d_4, d_5$ | 0.60 |
| $d_2 \rightarrow d_3$ | **0.75** |
| $d_2 \rightarrow d_3, d_4$ | 0.60 |
| $d_2 \rightarrow d_3, d_5$ | 0.60 |
| $d_2 \rightarrow d_3, d_4, d_5$ | 0.60 |

each rule's redundancy against the closest and more robust rule to it in an association basis defined by minimum support and confidence. Rules with a specific redundancy can be selected employing the imposition of a confidence boost threshold $\beta$. This parameter acts as a minimum threshold of redundancy that must be attained by all rules to be declared as irredundant. Rules below $\beta$ are redundant respect to the ones above it.

**Definition 6** (*Confidence Boost*)**.** The confidence boost of rule $r_0 : X_0 \rightarrow Y_0$, given a set of rules $r_i : X_i \rightarrow Y_i$, a set of trajectories $T$, and assuming $X \cap Y = \emptyset$, is:

$$cb(r_0, T) = \frac{c(r_0, T)}{max\left\{c(r_i, T) | r_i \neq r_0, s(\{X_i \cup Y_i\}, T) \geq \tau, X_i \subseteq X_0, Y_0 \subseteq Y_i\right\}}$$

By convention, if the set in the denominator is empty, the confidence boost reaches infinite; while if the set in the numerator is empty, the confidence boost will be zero. See results Section 3.4 where this situation happens. CauRuler selects rules with a low level of redundancy, therefore with a confidence boost over the minimum threshold of redundancy ($cb(r, T) > \beta$).

**Example 3.** Suppose we have a set of rules as the one introduced in Table 3. In this dataset, rule $d_1 \rightarrow d_2, d_3$ has a confidence 0.8. To measure its confidence boost one must consider all rules $X' \rightarrow Y'$ with $X' \subseteq \{d_1\}$ and $\{d_2, d_3\} \subseteq Y'$ listed in Table 3: One can see that the maximum confidence among them is 0.75, attained by $\{d_2, d_3\}$.

Then $cb(d_1 \rightarrow \{d_2, d_3\}, T) = 0.8/0.75 = 1.06667$. If we impose a threshold $\beta = 1.05$, rule $d_1 \rightarrow d_1, d_2$ will be considered a non-redundant rule.

### 2.3. Causality association detection

The algorithm evaluates the causality of the set of associations $R = \{r_1, \ldots, r_W\}$, with the aim of obtaining the set of causal rules $R^c = \{r_1^c, \ldots, r_{W^c}^c\} \subseteq R$ given the trajectories of patients $T_b = \{T_{p_1}^b, \ldots, T_{p_M}^b\}$. The final aim is to determine if there is a significant difference in the effects represented in patients undergoing the consequent of a causal rule ($Y$) between the ones affected and non-affected by diagnoses in the antecedent of the rule ($X$). The methodology consists of the steps outlined in the following subsections: creation of cohorts, controlling for confounder variables, fair dataset generation, association causality study and the generalization of causal rules. This procedure, summarized in Fig. 1, is applied to every non-redundant association rule.

#### 2.3.1. Cohort generation

To evaluate the causality of an association rule $r : X \rightarrow Y$ the binary trajectory dataset $T^b$ is separated into the *control* (healthy people) and *exposition* cohorts (people who suffered the health conditions in $X$). Given a rule $r : X \rightarrow Y$, where $X, Y \subseteq D$, *exposition cohort* and *control cohort* are defined as:

1. **Exposition cohort** ($X^e$)**:** the set of binary trajectories $T^b$ that contains diagnostics in $X$, that is, $X^e = \{t_{p_i}^b | \forall d_j^{p_i} \in X, \exists b_k^{p_i} \in t_{p_i}^b$ SUCH THAT $pos(d_j^{p_i}, D) = k, b_k^{p_i} = 1\}$.
2. **Control cohort** ($X^c$)**:** the set of binary trajectories $t_{p_i}^b$ that do not contains diagnostics in $X$, that is, $X^c = \{t_{p_i}^b | \forall d_j^{p_i} \in X, \exists b_k^{p_i} \in t_{p_i}^b$ SUCH THAT $pos(d_j^{p_i}, D) = k, b_k^{p_i} = 0\}$.

It holds that $X^c = T^b - X^e$.
For more information, see algorithm 1 in Appendix B.

#### 2.3.2. Confounders variables controlling

To conduct the causality study it is necessary to identify the confounding variables set $C$. Confounding bias occurs when a variable influences both who is selected for the exposition and the consequent variables in $Y$. Confounders may be known or merely suspected; acting as a 'lurking third variable' [12]. Thus, the apparent measure between an antecedent $X$ and consequent $Y$ may actually be due to another factor.

Confounder controlling is a difficult task. Overcontrolling could end up conditioning on a common cause, therefore not detecting the causal association. On the other hand, not controlling for the correct variables could end up retrieving spurious causality associations [11]. Leaving a confounder uncontrolled may lead to false discoveries while controlling for a confounder can cause the loss of true discoveries. In our framework, the heuristics promote the loss of some true discoveries in contrast to not getting false discoveries. The algorithm tends to overcontrol (control for a high number of different variables) in return to losing some true causal association, as in the medical field it is normally better to retrieve high-reliability results. Therefore, the algorithm does not detect the full causal map but detects a close partial one with high reliability on the causal inference procedure. Given an association rule $r : X \rightarrow Y$ we would like to control all possible variables $B$ not present either on $X$ or $Y$. This is to say, ideally we would like to control as many variables as possible, to isolate the effect of the exposure to the causal agent over the final effect.

With the increasing number of controlled variables, it will become more difficult to find the pairs of individuals between both cohorts to conduct a matching in the causality study. Therefore, there exists a trade-off between the dimension of controlled variables and the number of individuals needed in each cohort. Once the $X^e$ and $X^c$ cohorts are obtained, confounding variables will be selected after the identification of the irrelevant and exclusive variable sets, related to the target rule $r : X \rightarrow Y$. Irrelevant and exclusive variable sets are defined to avoid the aforementioned tradeoff. Not controlling variables with a lower cardinality in the trajectory database or in both individual cohorts increments the number of matchings to conduct the causality study 7. To determine the controlled variable set $C$, we first must identify the irrelevant and exclusive variable sets based on a predefined support of a binary set.

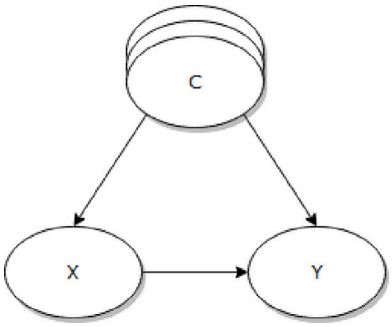


**Fig. 3.** Example causal map tested by CauRuler in each causality experiment. $X$ represent Left Hand-Side ($LHS$) set of rule $r : X \rightarrow Y$ while $Y$ represents Right Hand-Side ($RHS$) set. Confounders are represented by $C$ and may make an impact in both $X$ and $Y$ and producing spurious associations if not controlled.

**Definition 7** (*Support of a Binary Set*)**.** The support $s^b(Z^b, T^b)$ of a set of binary variables $Z^b \subseteq B$ regarding a set of binary trajectories $T^b$ is the cardinality of the set of clinical trajectories in $T^b$ containing $Z^b$: $s(Z^b, T^b) = |\{t^b_{p_i} \in T^b | Z^b \subseteq t^b_{p_i}\}|$.

**Definition 8** (*Irrelevant Variable*)**.** Given a set of binary trajectories $T^b$, each $t^b_{p_i} \in T^b$, and a rule $r : X \rightarrow Y$, a variable $b_j$ is irrelevant if it does not belong to either X or Y (i.e. $\nexists j \notin X \cup Y$) and it appears in an amount lower than $\epsilon_I$ in the overall set of trajectories $T^b$ (i.e. $s^b(\{b_j\}, T^b) < \epsilon_I$).[2]

Irrelevant variables $I$ (with $I \subset B$) do not have enough impact to justify the cost of controlling for them (are not sufficiently present in the dataset). $I$ is an approximation to improve the over-controlling behaviour of the algorithm. It is worth observing that there could be some diagnoses (not belonging to the rule r: $X \rightarrow Y$ under study) which could be irrelevant, as well as control variables. On the other hand, exclusive variables $E$ are identified as those variables not sufficiently present in one of the cohorts according to a given threshold $\epsilon_E$.

**Definition 9** (*Exclusive Variable*)**.** Given a cohort $X^k$, each $t^b \in X^k$, a variable $b_j$ is exclusive if it has informed values in an amount lower than $\epsilon$ in $X^k$ (i.e. $s(\{t^b | b_j = 1\}, X^k) \leq \epsilon_E$), and does not belong to either X or (i.e. $\nexists j \notin X \cup Y$).

If for example $\epsilon_E = \infty$ is set, it may happen that variable $b_j$ may be only present in one of the cohorts (a very rare health condition). Thus variable $b_j$ is marked as an exclusive variable, and added to the set of exclusive variables $E$.

The set of confounding variables is therefore determined by $C = B \setminus (I, E, X, Y)$ (see algorithm 2 in Appendix B.). Confounding variables enable the definition of the causal map regarding the rule $r : X \rightarrow Y$ under study (see Fig. 3).

The cohorts are revised according to $C$, $X'^e$ and $X'^c$, where in each trajectory irrelevant and exclusive variables have been removed. Therefore, the dimension of each revised binary trajectory is $|X \cup C \cup Y|$.

### 2.3.3. Fair dataset construction

A fair dataset $D_f$ is built using the *matching record pairs* concept. The aim is the acquisition of a fair dataset with the maximum number of pairs of individuals from each cohort that is as similar as possible between them. Using this approach we ensure the quantification of the effect variable, rather than the possible latent confounders.

[2] By convention, in the results it is used $\epsilon_I = \tau$; the threshold of the support used in the rule miner algorithm.

**Table 4**
Cohort example dataset.
*Source:* Adapted from [13].

| Cohort | Patient | $b_1$ (X) | $b_2$ | $b_4$ | $b_5$ | $b_6$ | $b_7$ | $b_3$ (Y) |
|---|---|---|---|---|---|---|---|---|
| $X^{fe}$ | 1 | **1** | 0 | 1 | 0 | 0 | 1 | **1** |
| | 2 | **1** | 0 | 1 | 0 | 1 | 0 | **1** |
| | 3 | **1** | 1 | 0 | 1 | 0 | 0 | **0** |
| | 4 | **1** | 1 | 0 | 0 | 0 | 1 | **1** |
| $X^{fc}$ | 5 | **0** | 0 | 1 | 0 | 0 | 1 | **0** |
| | 6 | **0** | 0 | 1 | 0 | 1 | 0 | **0** |
| | 7 | **0** | 1 | 0 | 1 | 0 | 0 | **0** |
| | 8 | **0** | 1 | 0 | 1 | 0 | 0 | **1** |

**Definition 10** (*Matched Record Pair*)**.** Given an association rule $r : X \rightarrow Y$, an exposure control cohort $X'^e$ and a control cohort $X'^c$, and a set of controlled variables $C$ derived from them, a matched record pair $\langle t'^e, t'^c \rangle$ ($t'^e \in X'^e$ and $t'^c \in X'^c$) is a pair of revised binary trajectories which share the same values for each variable $b'_j \in C$.

The resulting dataset $D^f$ contains $M'$ trajectories from $X'^e$ ($X^{fe} \subseteq X'^e$) and $M'$ trajectories from $X'^c$ ($X^{fc} \subseteq X'^e$), with $M' \leq |X'^e|, |X'^c|$; this is to say, $D_f = X^{fe} \cup X^{fc}$. For more detailed information, see algorithm 3 in Appendix B.

**Example 4** (*Fair Dataset Construction*)**.** Suppose a diagnosis set $D = \{d_1, d_2, d_3\}$, an association rule $X \rightarrow Y$, with $X = \{d_1\}$ and $Y = \{d_3\}$. Moreover, assume that set of binary diagnosis is $\{b_1, b_2, b_3, b_4, b_5, b_6, b_7\}$, with $b_1, \ldots, b_3$ representing to $d_1, d_2$, and $d_3$. Consistently, the controlled variable set is $C = \{b_2, b_4, b_5, b_6, b_7\}$. The available cohorts are shown in Table 4.

As it is possible to observe, patient 1 matches with patient 5, since the controlled variables have the same values for all of them ($b_2 = 0, b_4 = 1, b_5 = 0, b_6 = 0, b_7 = 1$). Analogously, matches $(2, 6), (3, 7)$ can be found. The final fair cohorts are $X^{fe} = (1, 2, 6)$ $X^{fc} = (5, 6, 7)$, and therefore $D_f = (1, 2, 3, 5, 6, 7)$. Each record can only be matched once, and appear once in the fair dataset ($D_f$). Therefore, it is possible to observe some members of $X'^e$ and $X'^c$ have no pairs and have been excluded in $D_f$. This is an important issue, as it will be latter seen, because the size of $D_f$ will determine the generation of the causal association rules.

### 2.3.4. Causality evaluation

The fair data set of a rule ($D_f$) helps us to simulate a controlled cohorts trial to test the hypothesis that the association rule ($r : X \rightarrow Y$) is causal. If we are controlling all possible confounder variables, we are demonstrating that the variation of the response variables ($Y$) is exclusively due to variables in $X$ (assuming all possible confounders are being controlled).

The causality of an association is evaluated through the lower confidence interval of its Odds Ratio, controlling for all possible confounder variables $C$.

**Definition 11** (*Odds Ratio of a Fair Dataset*)**.** Given an association rule $r : X \rightarrow Y$, and the fair dataset $D_f$ derived from it, with pairs of records from an exposure cohort $X^{fe}$, and records from a control cohort $X^{fc}$, the $OR$ of fair the dataset $D_f$ is defined as follows:

$$OR_{D_f}(r : X \rightarrow Y) = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}}$$

where:

- $n_{11}$ is the number of trajectories of patients with the diagnoses in $Y$ present in the exposure group. $n_{11} = |\{t | t \in X^{fe}, \forall d_i \in Y \exists b_j \in t,$ such that $pos(d_i, D) = j,$ and $b_j = 1\}|$
- $n_{12}$ is the number of trajectories of patients with none of the diagnoses in $Y$ present in the exposure group. $n_{12} = |\{t | t \in X^{fe},$ $\forall d_i \in Y \nexists b_j \in t,$ such that $pos(d_i, D) = j,$ and $b_j = 0\}|$

- $n_{21}$ is the number of trajectories of patients with the diagnoses in $Y$ present in the control group. $n_{21} = |\{t|t \in X^{fc}, \forall d_i \in Y \exists b_j \in t,$ such that $pos(d_i, D) = j,$ and $b_j = 1\}|$
- $n_{22}$ is the number of trajectories of patients with none of the diagnoses in $Y$ present in the control group. $n_{22} = |\{t|t \in X^{fc},$ $\forall d_i \in Y \nexists b_j \in t,$ such that $pos(d_i, D) = j,$ and $b_j = 0\}|$

Those associations fulfilling the minimum confidence threshold $\alpha$ (with an OR Confidence interval above 1) arise as causal associations (see algorithm 4 in Appendix B).

**Definition 12** (*Significant Odds Ratio*)**.** The confidence Interval $CI$ of the odds ratio of the rule $r : X \rightarrow Y$ on the given fair dataset $D_f$, $OR_{D_f}(r : X \rightarrow Y)$, is defined as:

$$exp(ln(OR_{D_f}(r : X \rightarrow Y))) \pm z' \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = [\alpha_-, \alpha_+]$$

Where $z'$ is a standard normal deviate corresponding to each level of confidence.

For convention, we will use $z'$=1.96 for 95% confidence. $\alpha_-$ and $\alpha_+$ are the lower and upper bounds respectively of an odds ratio at a confidence level. If $\alpha_- > 1$, the odds ratio is significantly higher than 1, hence $X \rightarrow Y$ is a causal association. This results in a highly trustable process since for small datasets, the $CI$ is wider, making it more difficult for a rule to be significantly associated.

#### 2.3.5. Causal rule generalization

Anti-monotone properties are the main foundation for efficient rule mining. Apriori rule miners such as the one proposed by Agrawal et al. are based on those properties to reach high efficiency by pruning itemsets that for sure will not surpass the demanded thresholds [33].

**Definition 13** (*Anti-Monotone Property*)**.** Any measure $f$ possesses the anti-monotone property if for every itemset $X$ that is proper subset of itemset $Y$, i.e. $X \subset Y$, we have $f(Y) \leq f(X)$.

Definition 13 states that if an itemset is infrequent ($s(X) \leq \tau$), then all of its supersets must also be infrequent ($s(Y) \leq \tau$ for all $X \subset Y$). This strategy is used to trim the exponential search space of all possible rules in a smaller space of rules which fit the thresholds [10] (see example 1 in figure 2).

This definition is exploited in CauRuler to obtain all more general rules of a causal rule. For example, if the relationship (*obesity* → *diabetes*) holds to be causal, then it is certain that this relationship holds for both sexes if properly controlled. Therefore the rule is redundant to rules: (*obesity* + *sex=male* → *diabetes*) and (*obesity* + *sex=female* → *diabetes*).

Therefore, we analyse causal rules, with $|X| > 2$ (named complex causal rules) looking for the following conditions:

(a) Specific causal associations of an irredundant causal association. $X$ is formed by both causal and non-causal elements of $Y$. Non-causal elements add specificity to an already causal association. In the next example, $Z20.828$ is adding specificity to an already causal association.
*[N18.9: Chronic kidney disease, Z20.828: Contact with and exposure to other communicable diseases.] → [I10: Essential (primary) hypertension]*

(b) Complex causal associations where each element in $X$ individually represents a cause of $Y$. For example,
*[E11.9: Type 2 diabetes mellitus, I45.0: Right fascicular block] → [I10: Essential (primary) hypertension]*
This rule could be split into (E11.9 → I10) and (I45.0 → I10)

(c) Complex causal associations where all elements of $X$ individually do not represent a cause of $Y$ but do so when acting as an ensemble.

**Table 5**
Dataset features: Number of patients, coverage percentage in the dataset, length of the diagnostic set, diagnoses occurrences (word length of the database).

| | Patients | % ind. | $\|D\|$ | diag occur. |
|---|---|---|---|---|
| All ages | 352,440 | (100%) | 3,994 | 3,489,948 |
| Child | 43,514 | (12%) | 3,076 | 336,211 |
| Adolescent | 28,136 | (8%) | 2,704 | 162,563 |
| Primary adulthood | 75,671 | (22%) | 3,356 | 536,466 |
| Late adulthood | 100,008 | (28%) | 3,538 | 899,534 |
| Elderly | 102,680 | (29%) | 3,731 | 1,539,912 |

In (*a*), all elements in $X$ must be evaluated separately as a cause of $Y$ to discover antimonotonic causal associations (the general case). This is to say, each permutation of diagnoses in exposure set $X' \subseteq X$ of a complex rule is transformed into an association rule of the form $r' : X' \rightarrow Y$ for a later evaluation of its causality. This way, from the causal association rules $R^c = \{r_1, \ldots, r_{W^c}\}$, is obtained $R^g = \{r_1^g, \ldots, r_{W^g}^g\}$ general rules, with $W^c \leq W^g$. A final step evaluates in which of the three aforementioned cases falls each rule, and stores the correct one.

### 2.4. Dataset

CauRuler has been implemented using Python and tested on a Primary Care Services Information Technologies System database of the Health region of Central Catalonia (Catalonia, Spain) belonging to the Catalan Institute of Health (ICS), the main primary healthcare provider in Catalunya.

The database contains 3,555,799 visits to the primary care services between 2019 and 2020 (a period covering most of the COVID-19 pandemic). The dataset is comprised of both face-to-face visits (to primary health systems and home visits) and non-face-to-face (telephone and teleconsultations) corresponding to 376,486 individuals (among the 404,245 reference population in this health region).

Each observation contains both "active diagnosis" (diagnoses that are active at the moment of the visit, mostly chronic), "visit diagnosis" (diagnosis arising from the consultation) and variables defining the patient such as age, sex and level of healthcare coverage. Visits are preprocessed to construct the clinical trajectories dataframe and the binary clinical trajectories dataframe. Only visit diagnoses are used to construct the trajectories since active diagnoses do not have the coding date, thus the time for the patient undergoing the disease is not known. There are up to 3,994 diagnoses according to ICD-10 ontology [27]. Diagnosis coding is prone to bias effects made by healthcare professionals (the COVID-19 outbreak has also put a strain on the healthcare system, leading to a likely increase in the rate of codification errors).

Patient ages have been discretized according to the World Health Organization (WHO) classification: *Child (0–11), adolescent (12–19), primary adulthood (20–40), late adulthood (41–60) and elderly (61-∞)*. Age distributions can be found in Table 5.

### 2.5. Experimental set up

The experimental setup tests the hypothesis that the CauRuler algorithm is better at detecting real causal associations by testing fewer association rules than in previous works. The result is an approximation to the complete causal map of the dataframe. By using high demanding support, confidence and confidence boost thresholds, the algorithm has to retrieve relevant (previously known or plausible) medical causal relations to prove effectiveness. Besides, a group of medical experts with different backgrounds evaluated the output of the different algorithms to prove their effectiveness. On the other hand, a study of the output dimension is held to prove the efficiency concerning another proposition.

**Table 6**
Number of causality associations (n) and percentage of causes among the total number of associations rules. All experiments use as support threshold $\tau = 0.0012$. $\gamma$ is the confidence parameter and $m$ and $\beta$ are pruning parameters of the different algorithms.

| Method | $\gamma$ | | | | | |
|---|---|---|---|---|---|---|
| | 0.6 | | 0.7 | | 0.8 | |
| | *n* | % | *n* | % | *n* | % |
| Apriori-OR($m = 1$) | 44 | 0.005 | 29 | 0.004 | 5 | 0.0017 |
| Apriori-OR($m = 5$) | 37 | 0.006 | 26 | 0.005 | 5 | 0.0017 |
| Apriori-OR($m = 10$) | 11 | 0.005 | 10 | 0.005 | 5 | 0.0017 |
| Apriori-OR($m = 15$) | 2 | 0.002 | 1 | 0.001 | 5 | 0.0028 |
| Apriori-OR($m = 20$) | 2 | 0.007 | 1 | 0.003 | 0 | 0.0000 |
| CauRuler($\beta = 0.0$) | 46 | 0.006 | 30 | 0.006 | 5 | 0.0030 |
| CauRuler($\beta = 0.1$) | 42 | 0.012 | 27 | 0.015 | 3 | 0.0070 |
| CauRuler($\beta = 0.2$) | 28 | 0.018 | 15 | 0.032 | 2 | 0.0900 |
| CauRuler($\beta = 0.3$) | 11 | 0.030 | 3 | 0.029 | 0 | 0.0000 |
| CauRuler($\beta = 0.4$) | 9 | 0.320 | 2 | 0.200 | 0 | 0.0000 |

**Table 7**
Mean number of diagnoses taking part in the rule (as LHS or RHS). *R: association rules; $R^c$: Causal rules.*. All experiments use as support threshold $\tau = 0.0012$. $\gamma$ is the confidence parameter and $m$ and $\beta$ are pruning parameters of the different algorithms.

| Method | $\gamma$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.6 | | 0.7 | | 0.8 | | 0.9 | |
| | R | $R^c$ | R | $R^c$ | R | $R^c$ | R | $R^c$ |
| Apriori-OR($m = 1$) | 3.85 | 2.82 | 4.01 | 3.03 | 4.51 | 3.40 | 5.24 | 0.00 |
| Apriori-OR($m = 5$) | 4.08 | 2.86 | 4.18 | 3.03 | 4.51 | 3.40 | 5.24 | 0.00 |
| Apriori-OR($m = 10$) | 4.55 | 3.09 | 4.45 | 3.20 | 4.58 | 3.40 | 5.24 | 0.00 |
| Apriori-OR($m = 15$) | 4.85 | 2.50 | 4.82 | 3.00 | 4.74 | 0.00 | 5.24 | 0.00 |
| Apriori-OR($m = 20$) | 5.05 | 2.50 | 5.08 | 3.00 | 4.87 | 0.00 | 5.24 | 0.00 |
| CauRuler($\beta = 0.0$) | 4.03 | 2.82 | 4.17 | 3.03 | 4.52 | 3.40 | 5.20 | 0.00 |
| CauRuler($\beta = 0.1$) | 3.57 | 2.73 | 3.66 | 2.92 | 3.92 | 3.00 | 3.00 | 0.00 |
| CauRuler($\beta = 0.2$) | 3.21 | 2.61 | 3.05 | 2.86 | 3.04 | 3.00 | 0.00 | 0.00 |
| CauRuler($\beta = 0.3$) | 3.06 | 2.18 | 2.96 | 2.33 | 3.00 | 0.00 | 0.00 | 0.00 |
| CauRuler($\beta = 0.4$) | 2.21 | 2.00 | 2.40 | 2.00 | 0.00 | 0.00 | 0.00 | 0.00 |

#### 2.5.1. Experimental scenario

To validate the CauRuler algorithm, different experimental scenarios were designed to compare it with the Apriori-OR algorithm [13], the causal rule mining approach more similar to CauRuler and being the state of the art in causality analysis using ARMs (see Section 2). All methods use the same parametrization to compare the results: *support* $\tau = 0.0012$, and *confidence* ($\gamma$); 0.6,0.7,0.8,0.9. Support is experimentally low-adjusted to target those more frequent rules. Apriori-OR adds the ($m$) parameter (minimum OR ratio to be considered as an association rule); 1,5,10,15,20. CauRuler employs parameters $\beta = 0.3$, $\epsilon = 10$, $\alpha = 1.0$. When evaluating the effect of the redundancy notion ($\beta$) takes values; $0.0, 0.1, 0.2, 0.3, 0.4$.

The goal of this section is to prove the algorithm's effectiveness at detecting reliable causal associations. For this reason, the parametrization used is very restrictive, in order to retrieve previously known causal associations that can be validated by medical experts. Conversely, CauRuler should prove efficiency by detecting as many causal relations as Apriori-OR by testing a smaller basis of association rules, thus reducing the number of causal inference trials.

## 3. Results

The causal associations obtained by the algorithms are collapsed in a causal map that is evaluated by medical experts. This strategy is easy to understand and visually attractive while emphasizing the multi-causal behaviour of some clinical conditions. The experts evaluated the plausibility of the associations targeted by the models.

The output dimension comparison of the algorithms is evaluated in terms of: number of causal association rules learnt by the algorithm, ratio of causal rules in the association rule basis, complexity of the causal rules learnt.

### 3.1. Causality findings

Table 6 summarizes the results of the number of causal associations obtained by the different algorithms. Results for $\gamma = 0.9$ are not provided since there are no causal rules found for this high demanding confidence threshold.

CauRuler algorithm can reduce the number of associations without losing much of the causal associations. This is to say, for $\beta = 0.0$ (without pruning) 46 causal associations were obtained, representing a 0.006% of all found association rules. With the increase in $\beta$, the % of causal associations found concerning the total number of associations rules is increased. The results are exacerbated for less demanding confidence thresholds. Apriori-OR algorithm does not target as many causal associations as the CauRuler algorithm.

While the ratio of causal rules over the total basis of associations is constantly increasing when boosting the $\beta$ parameter of CauRuler algorithm, increasing the $m$ parameter of Apriori-OR algorithm has a reverse effect. This behaviour suggests that while both algorithms perform better at reducing the basis of association rules, CauRuler algorithm is better at keeping the causal relations within the basis.

Deepening on the structure of the causal association rules, Table 7 summarizes the complexity of the different association rules concerning the number of diagnoses taking part in them. Causal rules, on average, present far fewer diagnoses than association rules. CauRuler can generalize the associations and, therefore, reduce their complexity, which might be useful for the interpretation of the results by humans. In contrast, Apriori-OR algorithm does not reduce the length of the associations. Therefore, it is not simplifying the rules like CauRuler.

### 3.2. Generalization analysis

All the causal associations found by CauRuler and Apriori-OR were subjected to the generalization analysis as explained in 2.3.5. From all combination within each rule, 14 general causal rules were obtained. No examples of type ($c$) complex causal association have been found. However multiple references to cases $(a, b)$ and simple causal associations (formed by two diagnoses) are found in the datasets. Table 8 summarizes the general causal associations found from the discovered ones sorted by lower confidence interval of the Odds Ratio.

Table 9 shows the number of associations referencing each general causal association described in Table 8. The experiment takes as reference the standard parametrization ($support = 0.0012, confidence = 0.6$) and the different values of $m$ and $\beta$ parameters used in previous examples. When increasing the $\beta$ parameter in CauRuler algorithm, the number of rules subjected to the causality study ($nR$) drops significantly. Similar behaviour, not as great as in the previous case, is seen when increasing $m$ parameter in Apriori-OR. Nevertheless, in the former case, the number of causal rules referencing each general causal rule ($nC$) and the number of general causal rules ($GC$) drops more slowly, while in the latter the dropping of number of association rules ($nR$) is linked to a dropping in ($nC$) and ($GC$). For example, applying a point increase in the confidence boost ($\beta = 0.1$) implies the reduction of 54% of causality studies while still detecting 55 causal associations and all the general causal rules (14, 100%); with a high demanding confidence boost threshold ($\beta = 0.4$) the reduction in causality studies reach 99.99% while still detecting 64% of general causal rules. In the case of the Apriori-OR, the results are worse. An increase in 5 points in the $OR$ threshold ($m = 5$) implies a reduction of 33% of causality rule studies detecting 37 causal rules and 71% of general causal associations. With a large increase in the threshold ($m = 20$) the reduction of causality studies reaches a reduction of 99.96% in causality studies while only targeting 7% of the general causal rules; far from the 64% of CauRuler.

The $CauRuler$ algorithm can reduce the basis of rules using a notion of redundancy without greatly compromising the inference of causality.

**Table 8**
Obtained general causal rules.

| ID | RULE |
|---|---|
| 1 | *['E78.00' = Pure hypercholesterolemia, unspecified] → ['I10' = Essential (primary) hypertension]* |
| 2 | *['N18.9' = Chronic kidney disease, unspecified] → ['I10' = Essential (primary) hypertension]* |
| 3 | *['E79.0' = Hyperuricemia without signs of inflammatory arthritis and tophaceous disease] → ['I10' = Essential (primary) hypertension]* |
| 4 | *['I45.0' = Right fascicular block] → ['I10' = Essential (primary) hypertension]* |
| 5 | *['E78.2' = Mixed hyperlipidemia] → ['I10' = Essential (primary) hypertension]* |
| 6 | *['E11.9' = Type 2 diabetes mellitus without complications] → ['I10' = Essential (primary) hypertension]* |
| 7 | *['E78.1' = Pure hyperglyceridemia] → ['I10' = Essential (primary) hypertension]* |
| 8 | *['R73.9' = Hyperglycemia, unspecified] → ['I10' = Essential (primary) hypertension]* |
| 9 | *['I10' = Essential (primary) hypertension] → ['E11.9' = Type 2 diabetes mellitus without complications]* |
| 10 | *['E11.319' = Type 2 diabetes mellitus with unspecified diabetic retinopathy without macular edema] → ['E11.9' = Type 2 diabetes mellitus without complications]* |
| 11 | *['R80.9' = Proteinuria, unspecified] → ['E11.9' = Type 2 diabetes mellitus without complications]* |
| 12 | *['U07.1' = COVID-19] → ['Z20.828' = Contact with and (suspected) exposure to other viral communicable diseases]* |
| 13 | *['R50.9' = Fever, unspecified] → ['Z20.828' = Contact with and (suspected) exposure to other viral communicable diseases]* |
| 14 | *['E21.3' = Hyperparathyroidism, unspecified] → ['N18.9' = Chronic kidney disease, unspecified]* |

Moreover, when increasing the redundancy, the rate of causal rules found per each general causal rule $nC/GC$ is reduced. Thus, the algorithm can directly retrieve general causal associations when increasing demanding confidence boost thresholds.

### 3.3. Redundancy reduction analysis

Table 10 provides detailed results concerning the use of the redundancy-based pruning method. A point increase in $\beta$ in CauRuler, with a low confidence threshold ($\gamma = 0.6$) implies a reduction of 54% of rules ($100 - \frac{3577}{7732}$); with a high ($\gamma = 0.9$) up to a 97% ($100 - \frac{4}{119}$). This reduction trend increases with higher values of $\beta$. Balcazar suggests using a *confidence boost threshold* $\beta = 1 - \gamma$ to achieve a proper result [19]. In the case of the Apriori-OR, the reduction trends work in the inverse direction: increasing when the confidence threshold increases. For a low min $m$ value of 5, and a low confidence threshold, the reduction achieved is 33%, and with a high *gamma* (= 0.9) no reduction is achieved.

### 3.4. Control of confounding variables analysis

Many of the associations obtained give inconclusive results in the causality study. This is due to an insufficient number of participants in the fair dataset to compute the Odds Ratio. Table 11 summarizes the average number of patients forming the generated fair datasets. It is noticeable that only 29% of the rules dispose of a sufficient number of patients (56 on average) to learn a causality result. In the remainder of the cases, 55% is due to a lack of coincidences concerning the effects of the association rules ($Y$ in $X \rightarrow Y$) in the control cohort, and a 16% is due to the inverse situation.

The number of participants in the fair dataset is conditioned by the controlled variable set $C$: CauRuler algorithm aims to impose as many control variables as possible, which limits the number of pairs of participants in the cohorts that can be matched to enter the fair dataset. By Definition 11, there are two possible situations in which one of the cohorts could end up with few or no individuals. In an extreme situation:

1. All cases forming the fair data set of a rule $r$ present the effect diagnoses $Y$ for the cohort $X^e$ while for $X^c$ cohort all cases does not present $Y$. Hence, the $OR$ numerator is 1 while its denominator is 0, resulting in an infinite value.
2. All cases forming the fair data set of a rule $r$ does not present the diagnoses $Y$ for the cohort $X^e$; while for $X^c$ cohort all cases present $Y$. Hence, the numerator of the $OR$ is 0 while its denominator is 1, resulting in a 0 value.

Only 29% of the associations present a conclusive result, summarized in Table 11. On average, rules with a conclusive result have fair datasets with 56 participants. However, the variability is high. On average, causal associations present 460 participants (see two last rows of Table 11). Those results are mainly due to the fact that the complexity of confirmed causal associations is lower when compared to those remaining unconfirmed. Causal rules associate 2.82 diagnoses on average while non-causal ones associate 3.35. Fewer diagnoses imply it is easier to find patients to enter the cohorts, and therefore, bigger fair datasets.

Increasing threshold values $\epsilon_I, \epsilon_E$ generates bigger $E, I$ subsets of diagnoses, which turn to entail a reduction in the dimension of the controlled variable set $C$. Employing it, the restriction regarding the record matching is lower and the causality study presents bigger fair datasets, reducing the number of inconclusive causality studies. Alternatively, it is always possible to increase the number of patients in the raw database.

**Table 9**
Number of subrules referencing each anti-monotone causal association from Table 8 with different algorithms and parametrizations. Table 1: Subrules obtained with $CB$ algorithm and boost parameter. Table2: Subrules obtained for $AOR$ algorithm and $OR$ parameter. Results obtained from computed datasets for $\gamma = 0.6$; *n rules:* Number of rules tested; *n causes:* Number of causal rules referenced; *n subcauses:* Number of subrules referencing each causal rule; *% rules:* percentage of causal rules found; *s/r:* Mean number of subrules found per causal rule.

| Method | General rules ID | | | | | | | | | | | | | | nR | nC | nGC | % GC | C/GC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | | | | | |
| Apriori-OR($m = 1$) | 9 | 6 | 1 | 2 | 2 | 16 | 2 | 1 | 0 | 1 | 1 | 10 | 1 | 2 | 8173 | 54 | 13 | 92% | 4.15 |
| Apriori-OR($m = 5$) | 7 | 4 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 1 | 1 | 5 | 1 | 2 | 5461 | 37 | 10 | 71% | 3.70 |
| Apriori-OR($m = 10$) | 3 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 2181 | 10 | 5 | 35% | 2.00 |
| Apriori-OR($m = 15$) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 781 | 0 | 1 | 7% | 2.00 |
| Apriori-OR($m = 20$) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 280 | 0 | 1 | 7% | 2.00 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | | | | | |
| CauRuler($\beta = 0.0$) | 9 | 6 | 1 | 2 | 2 | 16 | 2 | 1 | 0 | 1 | 1 | 10 | 1 | 2 | 7732 | 55 | 14 | 100% | 3.92 |
| CauRuler($\beta = 0.1$) | 6 | 4 | 1 | 1 | 1 | 13 | 2 | 1 | 1 | 1 | 1 | 10 | 1 | 2 | 3577 | 45 | 14 | 100% | 3.21 |
| CauRuler($\beta = 0.2$) | 4 | 2 | 1 | 1 | 1 | 8 | 1 | 1 | 1 | 1 | 0 | 7 | 1 | 1 | 1512 | 30 | 13 | 92% | 2.30 |
| CauRuler($\beta = 0.3$) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 356 | 11 | 11 | 78% | 1 |
| CauRuler($\beta = 0.4$) | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 28 | 9 | 9 | 64% | 1 |

**Table 10**
Number of rules obtained by each method according to the different parameters: $n$ is the amount of discovered rules; $m$ is the minim OR ratio for considering redundant rules in Apriori-OR; $\beta$ the threshold to prune redundant rules in Cauruler; $\gamma$ the confidence threshold. All experiments use as support threshold $\tau = 0.0012$.

| Method | $m$ | $\beta$ | $\gamma$ | | | |
|---|---|---|---|---|---|---|
| | | | 0.6 | 0.7 | 0.8 | 0.9 |
| Apriori | – | – | 8,173 | 5,923 | 2,801 | 231 |
| Apriori-OR | 1 | – | 8,173 | 5,923 | 2,801 | 231 |
| Apriori-OR | 5 | – | 5,461 | 5,019 | 2,801 | 231 |
| Apriori-OR | 10 | – | 2,181 | 1,952 | 2,801 | 231 |
| Apriori-OR | 15 | | 781 | 728 | 1,778 | 231 |
| Apriori-OR | 20 | – | 280 | 298 | 279 | 138 |
| CauRuler | – | 0.0 | 7,732 | 4,714 | 1,932 | 119 |
| CauRuler | – | 0.1 | 3,577 | 1,796 | 460 | 4 |
| CauRuler | – | 0.2 | 1,512 | 460 | 22 | 0 |
| CauRuler | – | 0.3 | 356 | 101 | 4 | 0 |
| CauRuler | – | 0.4 | 28 | 10 | 0 | 0 |

**Table 11**
Analysis of the factors that regulate the generation of causal association rules. Parameters for the setting: $\gamma = 0.6$ and $\beta = 0.0$. Factors: $n$: number of rules; $|\vec{D}_f|$: mean number of participants taking part in the study of the rule; $|X \cup Y|$: mean number of diagnosis taking part in rules analysis. "Causal True" and "Causal False" are subsets of rules from the "Computable" set. Computed rules can be either causal or non-causal, while for non-computed ("OR∞", "OR0") the causality is unknown.

| Condition | n (%) | $\lvert\vec{D}_f\rvert$ | $\lvert\bar{C}\rvert$(%) | $\lvert X \cup Y\rvert$ |
|---|---|---|---|---|
| OR∞ | 4268 (0.55) | 4 | 1510 (0.38) | 4.44 |
| Computable | 2240 (0.29) | 56 | 1389 (0.35) | 3.34 |
| OR0 | 1224 (0.16) | 15 | 1505 (0.37) | 3.84 |
| Causal True | 46 (0.02) | 460 | 906 (0.23) | 2.82 |
| Causal False | 2194 (0.98) | 48 | 1399 (0.35) | 3.35 |

### 3.5. Medical evaluation analysis

To determine the clinical validity of the results, two causal maps are defined using as input the top-10 causal associations obtained by CauRuler 4 and Apriori-OR algorithm. The associations are selected based on the lower confidence interval of the Odds Ratio. The experiment takes as reference the standard parametrization ($support = 0.0012, confidence = 0.6$), CauRuler parameter $\beta = 0.3$ and Apriori-OR parameter $m = 1$. Three different medical experts from different backgrounds (neurology and general medicine) whom do not known each other were selected to evaluate both causal maps. The causal associations were evaluated to judge previous knowledge by the expert, plausibility if not known or reason why they are not plausible if its the case. The experts did not know which algorithm generated the causal rule at the time of assessment. Table 12 summarizes the obtained results.

It is interesting to say that expert 2 stated that causal association number 6 (*Type 2 diabetes → Primary hypertension*) in Table 8, although being plausible (diabetes might result in vascular alteration inducing hypertension, as some studies point) could be also caused by common factors such as obesity, smoking habits or sedentarism. This statement sustains the cycle that can be observed between both diagnoses in causal map (Fig. 4). In fact, a causal map by definition is a DAG, but this is not the case regarding the aforementioned two diagnoses. This cycle structure suggests the existence of a lurking third variable causing both diagnoses that is not being controlled. Since the experiment is controlling for smoking, using data about the Body Mass Index (BMI) or sedentarism could be a good approach to test causality over those associations (there could be other variables also acting as confounders). The expert also suggested that association 3 (*Hyperuricemia → Primary hypertension*) could have a similar behaviour. The aforementioned results are interesting since prove that: the expert knowledge always will be needed to evaluate or reinforce the results (suggesting interesting confounders to induce better results) and that from the causal maps it is possible to point structures suggesting a spurious causal association.

All experts reached consensus on the causal associations that are erroneous in the Apriori-OR causal map. Both rules are complex (formed by more than 2 diagnoses) and are associating cardiovascular diseases with covid. Those associations are formed by a true causal path but add a third term bringing specificity and incorrectness to the model. It is true that chronic kidney disease causes hypertension both for people undergoing covid-19 and people with no covid infection. Those incorrect results are obtained since the incorrect diagnose (covid) has not enough strength to lower the OR of the true causal association, thus retrieving a monotonic association falling in a specific case of the real causal association. The CauRuler algorithm is able to side-step the problem by using the confidence boost, which generalizes the rules without falling in too specific causal scenarios.

The erroneous targeted rules are:

- *["Pure hypercholesterolemia", "Covid-19", "Type 2 diabetes mellitus"] → hypertension*
- *["Chronic kidney disease", "Covid-19"] → hypertension*

The evaluation made by the medical experts shows that CauRuler algorithm is better at detecting causal associations. Nevertheless, the algorithm will always be subjected to errors related with not properly controlling for confounder variables. Medical experts and external information can be a source for detecting the variables that can act as confounders in different associations. Causal maps can be used to detect erroneous causal paths.

### 3.6. Discussion

The dimension of the dataset used with a diagnosis set (up to 3,994 diagnoses) and the number of patients considered (up to 352,440), as well as the variable length of the trajectories, make it a complex dataset [34]. The results achieved prove the scalability and effectiveness of the algorithm at targeting real causal associations. The algorithm is able to target interesting causal associations avoiding the coding errors present in the database (the primary healthcare system was highly affected by the covid-19 outbreak, specially in 2020).

Regarding the causality identification, CauRuler has shown efficacy in detecting a more complex causal map than the one achieved by Apriori-OR employing a smaller association rule basis. Causal associations targeted by the model were previously known and have consensus in the medical community. By applying causal rule generalization, the proposed method is able to get rid of those relations that fall in too specific scenarios, which tend to induce erroneous causal rules. This behaviour meets what was expected for these high demanding thresholds of parameters and prove the effectiveness of the model. As stated by experts, some of the causal associations targeted by both algorithms might be caused by a lurking third variable not presents in the dataset, thus not being controlled. Causal discovery algorithms assume they are controlling all confounders, but this is highly improvable in some cases. For the time being, the thresholds used were highly demanding in order to retrieve previously known associations to validate the results. Softening the thresholds in future experimentation strategies should retrieve other associations that might prove interesting to evaluate. During the Covid-19 outbreak, specially in the first stages, COVID-19 was coded using three different diagnoses ($B34.2$: Coronavirus infection, unspecified, $Z20.828$:Contact with and (suspected) exposure to other viral communicable diseases, $U07.1$:COVID-19). There are causal associations relating those terms. Causality would suggest covid causes covid, but those relations are instead explained due to the high pressure which was involved in the health system during the outbreak.

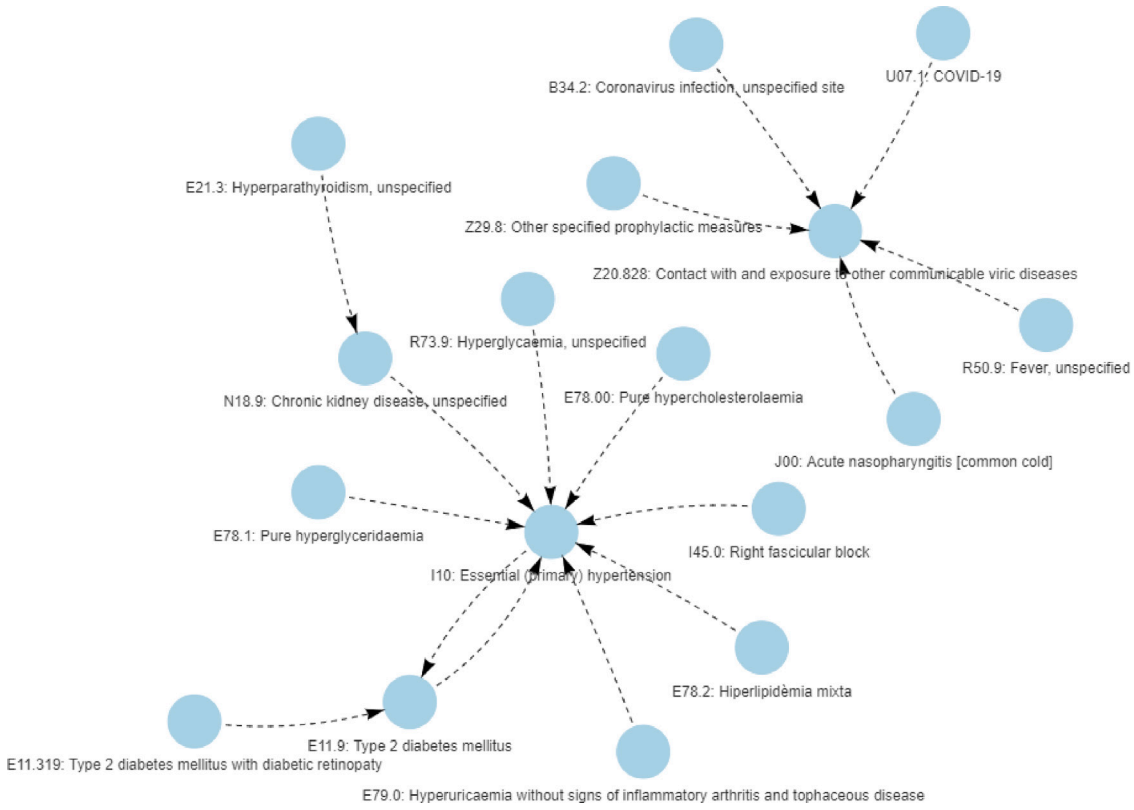


**Fig. 4.** Causal map obtained by CauRuler algorithm for $\beta = 0.0$ and $\gamma = 0.6$. General causal rules from Table 8. Results from ICS-Catalunya central Dataset over diagnoses between years 2019 and 2020 (covid-19 outbreak period).

**Table 12**
Medical evaluation of top-10 causal associations detected by CauRuler and Apriori-OR algorithm.

| | CauRuler algorithm | | | Apriori-OR algorithm | | |
|---|---|---|---|---|---|---|
| | Previously known | Plausible | Impossible | Previously known | Plausible | Impossible |
| Medical expert 1 | 100% | 0 | 0 | 80% | 0 | 20% |
| Medical expert 2 | 80% | 20% | 0 | 80% | 0 | 20% |
| Medical expert 3 | 100% | 0% | 0 | 80% | 0 | 20% |

CauRuler has shown a higher power than its predecessors to prune the number of possible association rules. The strength of the CauRuler algorithm is that it reduces the basis based on inclusion relations, and therefore, no information is lost (it only gets rid of redundant associations). Conversely, Apriori-OR prune rules based on the OR before applying any control of confounding variables, which could cause some information loss. The confidence boost proposition reduces the dimension of the associations basis without highly compromising the targeted number of causal associations. By using a smaller basis of rules, Cauruler is capable of targeting more Causal associations and retrieve more complex causal maps. This behaviour allows the mining of bigger databases, with a higher number of patients and variables.

## 4. Conclusions

This paper proposes an algorithm to target causal associations on a basis of non-redundant association rules. The method brings together irredundant rule mining and retrospective cohort study, controlling confounding variables, and using anti-monotone properties to learn a smaller, more general, set of causal rules. The carried out experimentation shows how CauRuler is capable of discovering partial causal maps from a complex clinical databases controlling a high number of confounder variables. The medical evaluation proves the reliability of the obtained associations.

In terms of efficiency evaluation, future work includes an exhaustive study of the computational times compared to other causal methods such as Bayesian networks. In terms of causality discovery, future work will include experimentation other databases containing more *clinic interest variables* that can act as confounders, such as socioeconomic, geo-clinical, analytical or pharmacological data.

Further analysis will include an exhaustive evaluation of causal associations accounting diagnoses and pharmacology. Conversely, studying specific cohorts of patients undergoing a particular condition or confounder (sex, age, smoking status...) will lead to causal association findings within the cohort, avoiding the effect of invisibilization when mining the whole population.

**Table 13**
CauRuler variables and parameters definition.

| Var | Description |
|---|---|
| $P = \{p_1, \dots, p_M\}$ | **Set of patients** |
| $p_i$ | Patient |
| $M$ | Number patients |
| $D = \langle d_1, \dots, d_N \rangle$ | **Tuple of diagnoses** |
| $d_i$ | Diagnostic (ICD-10 codification) |
| $N$ | Number of diagnoses |
| $V = \{v_1, \dots, v_K\}$ | **Set of variables of clinical interest** |
| $v_i$ | Variable |
| $K$ | Number of variables |
| $V^b = \{v^b_1, \dots, v^b_{K^b}\}$ | **Set of binary variables of clinical interest** |
| $v^b_i$ | Binary variable |
| $K^b$ | Number of clinical binary variables |
| $B = \{b_1, \dots, b_{N+K^b}\}$ | **Set if binary diagnoses and clinical binary variables** |
| $b_i$ | Binary variable |
| $T = \{t_{p_1}, \dots, t_{p_M}\}$ | **Set of Trajectories** |
| $t_{p_i} = \langle d^{p_i}_1, \dots, d^{p_i}_{n_{p_i}} \rangle$ | Transaction/Patient trajectory |
| $n_{p_i}$ | Number of diagnoses of patient $p_i$ |
| $T^b = \{t^b_{p_1}, \dots, t^b_{p_M}\}$ | **Set of Binary Transactions** |
| $t^b_{p_i} = \langle b^{p_i}_1, \dots, b^{p_i}_{N+K^b} \rangle$ | Binary trajectory of patient |
| $R = \{r_1, \dots, r_W\}$ | **Set of association rules** |
| $r_i : X_i \to Y_i$ | Association rule |
| $X_i, Y_i \subset D$ | Subsets of diagnoses forming rule |
| $W$ | Number of association rules |
| $LHS(r_i) = X_i$ | Left Hand Side of rule |
| $RHS(r_i) = Y_i$ | Right Hand Side of rule |
| $R^c = \{r^c_1, \dots, r^c_{W^c}\}$ | **Set of causal association rules** |
| $R^g = \{r^g_1, \dots, r^g_{W^g}\}$ | **Set of general causal association rules** |
| $X^e, X^c$ | Exposure and control cohorts |
| $X^{fe}, X^{fc}$ | Fair exposure and fair control cohorts |
| $D_f = X^{fe} \cup X^{fc}$ | Fair dataset |
| $I$ | Irredundant variable set |
| $E$ | Exclusive variable set |
| $C$ | Controlled variable set |
| $\tau$ | Support threshold |
| $\gamma$ | Confidence threshold |
| $\beta$ | Confidence boost threshold |
| $\epsilon_I, \epsilon_E$ | Irrelevant and Exclusion variables thresholds |
| $\alpha$ | Confidence interval threshold |

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This study was conducted with the support of the Secretary of Universities and Research of the Department of Business and Knowledge at the Generalitat de Catalunya, Spain (SGR 01125).

## Appendix A. Notation

Table 13 summarizes all variables and parameters used in the methodology.

## Appendix B. Code

Pseudocode representation of the different steps of the CauRuler algorithm. Source code is available in a private repository under query to the authors.

**Algorithm 1** COHORT GENERATION

- **Input:** binary trajectory dataset $T^b$; Set if binary diagnoses and clinical binary variables $B$; rule $r = X \to Y$.
- **Output:** exposition cohort $X^e$; Control cohort $X^c$.

**procedure** COHORTS OF RULE. ($Co(r)$)
  Let exposition and control cohorts $X^e = \emptyset, X^c = \emptyset$
  Let $X^b = \emptyset$ ▷ Binarization of X
  **for** each $b_j$ in $B$ **do**
    **if** $\exists d_k, pos(d_k \in X, D) = j$ **then**
      Add $b_j = 1$ to $X^b$
    **else**
      Add $b_j = 0$ to $X^b$
    **end if**
  **end for**
  **for** each $t^b_{p_i}$ in $T^b$ **do**. ▷ Cohort generation
    **if** $X^b \cap t^b_{p_i} = X^b$ **then**
      Add $t^b_{p_i}$ to $X^e$
    **else**
      Add $t^b_{p_i}$ to $X^c$
    **end if**
  **end for**
  Return $X^e, X^c$
**end procedure**

**Algorithm 2** CONFOUNDERS VARIABLE CONTROLLING

- **Input:** Binary trajectory dataset $T^b$; support threshold $\tau$; variable set $B$; Exposition cohort $X^e$ an control cohort $X^c$; exclusiveness threshold $\epsilon$.
- **Output:** Controlled variable set $C$.

**procedure** IRRELEVANT VARIABLE SET. ($I$)
  Let $I = \emptyset$
  **for** $b_i \in B \setminus X \cup Y$ **do**
    **if** $s(\{b_i\}, T^b) \le \epsilon_I$ **then**
      Add $d_i$ to $I$
    **end if**
  **end for**
  Return $I$
**end procedure**

**procedure** EXCLUSIVE VARIABLE SET. ($E$)
  Let $E = \emptyset$
  **for** cohort $X^k$ in $X^e, X^c$ **do**
    **for** $b_i \in B \setminus X \cup Y$ **do**
      **if** $s(\{b_i\}, X^k) \le \epsilon_E$ **then**
        Add $b_i$ to $E$
      **end if**
    **end for**
  **end for**
  Return $E$
**end procedure**

**procedure** CONTROLLED VARIABLE SET C. ($C$)
  Return $C = B \setminus (X, Y, I, E)$
**end procedure**

**Algorithm 3** FAIR DATASET CONSTRUCTION

- **Input:** Exposition Cohort $X^e$ and Control Cohort $X^c$, rule $r = X \rightarrow Y$, controlled variable set $C$.
- **Output:** Fair dataset $D_f(r)$ of rule $r : X \rightarrow Y$.

**procedure** FAIR DATASET OF RULE. $(D_f(r))$
  Let fair dataset $D_f = \emptyset$
  Assume $|X^e| < |X^c|$, if not swap.
  **for** each $t^b_{p_i}$ in $X^e$ **do**
    Let $t^b_{p_j} = match(t^b_{p_i}, X^c - D_f, C)$
    **if then**$t^b_{p_i}$
      add $t^b_{p_i}, t^b_{p_j}$ to $D_f$
    **end if**
  **end for**
  Return $D_f$
**end procedure**

**Algorithm 4** CAUSALITY EVALUATION

- **Input:** rule $r_i = X \rightarrow Y$; Fair dataset $D_f = X^{fe} \cup X^{fc}$ of rule $r$; standard normal derivate $z$.
- **Output:** boolean of causality.

**procedure** CAUSALITY EVALUATION FOR RULE $(caus(r))$
  $n11 = |\{t | t \in X^{fe}, \forall j \in Y \exists b_j \in t, b_j = 1\}|$
  $n12 = |\{t | t \in X^{fe}, \forall j \in Y \nexists b_j \in t, b_j = 0\}|$
  $n13 = |\{t | t \in X^{fc}, \forall j \in Y \exists b_j \in t, b_j = 1\}|$
  $n14 = |\{t | t \in X^{fc}, \forall j \in Y \nexists b_j \in t, b_j = 0\}|$
  Let $OR_{D_f} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$
  Let $\alpha_-$ be the minimum CI of $OR_{D_f}$
  **if** $\alpha_- \geq 1$ **then**
    Return TRUE
  **else**
    Return FALSE
  **end if**
**end procedure**

## References


[1] A. Morabia, Quality, originality, and significance of the 1939 "Tobacco consumption and lung carcinoma" article by Mueller, including translation of a section of the paper, Prev. Med. 55 (3) (2012) 171–177, http://dx.doi.org/10.1016/J.YPMED.2012.05.008.

[2] R. Riegelman, Contributory cause: Unnecessary and insufficient, Postgrad. Med. 66 (2) (1979) 177–179, http://dx.doi.org/10.1080/00325481.1979.11715231.

[3] G.F. Cooper, A simple constraint-based algorithm for efficiently mining observational databases for causal relationships, Data Min. Knowl. Discov. 1 (2) (1997) 203–224, http://dx.doi.org/10.1023/A:1009787925236.

[4] C. Bonell, J. Hargreaves, V. Strange, P. Pronyk, J. Porter, Should structural interventions be evaluated using RCTs? The case of HIV prevention, Soc. Sci. Med. 63 (5) (2006) 1135–1142, http://dx.doi.org/10.1016/j.socscimed.2006.03.026.

[5] J. Pearl, T.S. Verma, A theory of inferred causation, Stud. Logic Found. Math. 134 (1995) 789–811, http://dx.doi.org/10.1016/S0049-237X(06)80074-1.

[6] D. Heckerman, A Bayesian approach to learning causal networks, Adv. Decis. Anal. From Found. Appl. (2013) 202–220, http://dx.doi.org/10.1017/CBO9780511611308.012.

[7] D.M. Chickering, D. Heckerman, C. Meek, Large-sample learning of Bayesian networks is NP-hard, J. Mach. Learn. Res. 5 (2004) 1532–4435, http://dx.doi.org/10.5555/1005332.1044703.

[8] S. Mani, P. Spirtes, G.F. Cooper, A theoretical study of Y structures for causal discovery, in: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI '06, 1, AUAI Press, Arlington, Virginia, USA, 2006, pp. 314–323, http://dx.doi.org/10.5555/3020419.3020458.

[9] C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X.D. Koutsoukos, Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical, J. Mach. Learn. Res. 11 (2010) 171–234, http://dx.doi.org/10.5555/1756006.1756013.

[10] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, SIGMOD Rec. 22 (2) (1993) 207–216, http://dx.doi.org/10.1145/170036.170072.

[11] J. Pearl, Causality, second ed., Cambridge University Press, 2009, http://dx.doi.org/10.1017/CBO9780511803161.

[12] J. Pearl, D. Mackenzie, The book of why: The new science of cause and effect, 2019.

[13] J. Li, T.D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, Mining causal association rules, in: Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW, IEEE Computer Society, 2013, pp. 114–123, http://dx.doi.org/10.1109/ICDMW.2013.88.

[14] P. Yadav, M. Steinbach, M.R. Castro, P.J. Caraballo, V. Kumar, G. Simon, Frequent causal pattern mining: A computationally efficient framework for estimating bias-corrected effects, NIH Public Access, 2019, pp. 1981–1990, http://dx.doi.org/10.1109/BIGDATA47090.2019.9005977.

[15] M.J. Zaki, Scalable algorithms for association mining, IEEE Trans. Knowl. Data Eng. 12 (3) (2000) 372–390, http://dx.doi.org/10.1109/69.846291.

[16] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM SIGMOD Rec. 29 (2000) 1–12, http://dx.doi.org/10.1145/335191.335372.

[17] M. Luxenburger, Implications partielles dans un contexte, Math. Sci. Hum. 113 (1991) 35–55.

[18] M.J. Zaki, Mining non-redundant association rules, Data Min. Knowl. Discov. 9 (2004) 223–248, http://dx.doi.org/10.1023/B:DAMI.0000040429.96086.c7.

[19] J.L. Balcázar, Formal and computational properties of the confidence boost of association rules, ACM Trans. Knowl. Discov. Data 7 (4) (2013) http://dx.doi.org/10.1145/2541268.2541272.

[20] A. Fahmi, A. Macbrayne, E. Kyrimi, S. McLachlan, F. Humby, W. Marsh, C. Pitzalis, Causal Bayesian networks for medical diagnosis: A case study in Rheumatoid Arthritis, in: 2020 IEEE International Conference on Healthcare Informatics, ICHI, 1, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1–7, http://dx.doi.org/10.1109/ICHI48887.2020.9374327.

[21] F.L. Seixas, B. Zadrozny, J. Laks, A. Conci, D.C.M. Saade, A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment, Comput. Biol. Med. 51 (2014) 140–158, http://dx.doi.org/10.1016/J.COMPBIOMED.2014.04.010.

[22] D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, M. Perc, Saving human lives: What complexity science and information systems can contribute, J. Stat. Phys. 158 (2014) 735–781, http://dx.doi.org/10.1007/s10955-014-1024-9.

[23] N. Pombo, N. Garcia, K. Bousson, Classification techniques on computerized systems to predict and/or to detect Apnea: A systematic review, Comput. Methods Programs Biomed. 140 (2017) 265–274, http://dx.doi.org/10.1016/j.cmpb.2017.01.001.

[24] C.H. Wang, T.Y. Lee, K.C. Hui, M.H. Chung, Mental disorders and medical comorbidities: Association rule mining approach, Perspect. Psychiatr. Care 55 (3) (2019) 517–526, http://dx.doi.org/10.1111/ppc.12362.

[25] K.S. Lakshmi, G. Vadivu, Extracting association rules from medical health records using multi-criteria decision analysis, Procedia Comput. Sci. 115 (2017) 290–295, http://dx.doi.org/10.1016/j.procs.2017.09.137.

[26] S. Stilou, P. Bamidis, N. Maglaveras, C. Pappas, Mining association rules from clinical databases: An intelligent diagnostic process in healthcare, Stud. Health Technol. Inform. 84 (2) (2001) 1399–1403.

[27] WHO, International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). U.S, National Center for Health Statistics (NCHS), Department of Health & Human Services, 2019. URL http://www.cdc.gov/nchs/icd/icd10cm.htm.

[28] D. Allen, Automatic one-hot re-encoding for FPGAs, in: International Conference on Field-Programmable Logic and Applications, 705 LNCS, Springer, Berlin, Heidelberg, 1992, pp. 71–77, http://dx.doi.org/10.1007/3-540-57091-8_31.

[29] P. Wang, J. Xu, C. Wang, G. Zhang, H. Wang, Method of non-invasive parameters for predicting the probability of early in-hospital death of patients in intensive care unit, Biomed. Signal Process. Control 73 (2022) 103405, http://dx.doi.org/10.1016/J.BSPC.2021.103405.

[30] J. Pearl, From Bayesian networks to causal networks, Math. Model. Handl. Partial Knowl. Artif. Intell. (1995) 157–182, http://dx.doi.org/10.1007/978-1-4899-1424-8_9.

[31] D. Heckerman, Bayesian networks for data mining, Data Min. Knowl. Discov. 1 (1997) 79–119, http://dx.doi.org/10.1023/A:1009730122752.

[32] S. Nadkarni, P.P. Shenoy, Bayesian network approach to making inferences in causal maps, European J. Oper. Res. 128 (3) (2001) 479–498, http://dx.doi.org/10.1016/S0377-2217(99)00368-9.

[33] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, Fast discovery of association rules, Adv. Knowl. Discov. Data Min. (1996) 307–328.

[34] F. Lopez Segui, G. Hernandez Guillamet, H. Pifarré Arolas, F.X. Marin-Gomez, A. Ruiz Comellas, A.M. Ramirez Morros, C. Adroher Mas, J. Vidal-Alaball, Characterization and identification of variations in types of primary care visits before and during the COVID-19 pandemic in Catalonia: Big data analysis study, J Med Internet Res 23 (9) (2021) e29622, http://dx.doi.org/10.2196/29622.