diagnoses analysis through graph decomposition and association rules in the context of Covid-19

Final Master Thesis - Data Science

Master in Innovation and Research in Informatics
Facultat d'informàtica de Barcelona
Universitat Politècnica de Catalunya

Author:
Guillem Hernández Guillamet
guillemhg98@gmail.com

Supervisors:
Jose Luis Balcazar Navarro
jose.luis.balcazar@upc.edu

June 22, 2021

"This project aims to help health care professionals try to anticipate which initial pathologies may lead to the most prevalent chronic pathologies. This prevention, if effective, would lead to significant savings in health care resources, as these chronic pathologies represent a huge expense for the health care system."

Josep Vidal Alaball - *MD, coordinator innovation and research unit in primary care (Catalan ministry of health, Catalunya-central)*

"The digital revolution experienced by all the social environments has affected also the way of working of the medical services, gathering high amounts of data that could be useful for the clinical research. In the Catalan Public Healthcare system, the technique proposed has still way ahead, but asserts an initial formulation of a technique that could be useful for raising medical knowledge if exploited correctly in the future. Those data science techniques could change the way we understand the health of the population, improving their curing."

Francesc Lòpez Seguí - *Health Economist (Catalan Ministry of Health)*

Acknowledgements

First and foremost I would like to use this section to thank doctor/professor José Luis Balcázar for the opportunity he gave to me to work on this interesting project, giving my humble contribution to his already large and leading research on the topic, using and learning from his approaches and research.

I would also like to thank Francesc López Seguí and Josep Vidal Alaball for their constant support and dedication. Not only do I want to express my gratitude towards them for the unconditional support, but also for the countless opportunities in various challenging yet interesting projects they gave to me.

I cannot end this note without thanking my whole family, especially my parents and brother, who supported me throughout not only my academic career but also my life.

Last but not least, I need to mention my girlfriend Estel Figueras for the countless discussions on the topic that have significantly helped in improving and guiding me throughout this project, arising several important questions that have been key in finishing it. I would like to thank her not only for supporting me through academics but also in my everyday life, being patient and kind.

Abstract

Rule miners are unsupervised learning methods used to detect associations between items. These algorithms have been traditionally used in transactional datasets to synthesise significance associations between items. Extrapolating this behaviour to EHR data, the algorithms should be able to detect associations between diagnoses in a certain segment of the population, therefore suggesting relations of conditions prone to interest by the medical community.

This thesis provides an evaluation of a proposal of a rule mining algorithm to detect associations of diagnostses in medical trajectory databases of patients. The approach uses the notion of redundancy to solve the main issues of output size and validity traditionally suffered by rule miners by finding only the non-redundant significant associations. The yacaree program is able to use this approach reducing at the minimum level the needing of expertise by the end user. This thesis evaluates the validity of this technique in a high demanding medical dataset with respect to other rule miner approaches.

The procedure aims to state an initial proposal for mining EHR databases to detect between and within associations of diagnoses in segments of patients based on confounding factors age ans sex, with promising results. By imposing high-demanding thresholds the procedure is able to retrieve associations of diagnoses that although being evident suggest correctness of the approach. By softening the thresholds, one should be able to detect non-obvious associations prone to research. The method is tested in a database of visits during the covid-19 outbreak period to bring to light possible associations with the pandemic.

Using network visualizations, the ultimatre goal is making a primal formulation of a tool that can be easily interpreted by the medical community.

Two final research proposals are adressed. First the suggestion of a basic algorithm to detect morbidity groups of diagnoses. Second, the detection of directionality between diagnoses in rules to improve the visualization and suggest temporality, which turns to be very interesting from the medical perspective.

Keywords: Association rule miners, confidence, EHR, redundancy, network, algorithms.

Index

List of Figures

Network html repositories

1. Algorithm comparison plots
2. Item frequency plots
3. Trajectories Sizes
4. RAW, SEX, AGE yearly rule networks
5. RAW, SEX, AGE yearly directionality networks
6. Hierarchical dendograms morbidity detection
7. Community morbidity detection plots
8. Community morbidity rules networks
9. RAW, SEX, AGE year comparison rule networks
10. RAW, SEX, AGE year comparison directionality networks

List of Tables

Chapter 1

Introduction

The following chapter aims to put the basis and contextualize the algorithms and technical part of association rule miners that have been used in this thesis along with the research question that has been tackled. The main objective is to evaluate the algorithm into a real-case dataset in order not to only improve the knowledge of the medical field, yet proposing a reliable technique to improve medical decision-making. Moreover, an initial justification of the objectives to tackle will be held in this chapter.

1.1 State of the art

Association rule mining is a family of approaches intending to discover interesting relations between items in large transactional datasets by identifying frequent item-sets called association rules of the form "if-then". In [Agrawal et al., 1996] a first formal definition of an algorithm for association rule mining in large databases is proposed. To contextualize the algorithms, authors traditionally use a "supermarket database" comprising transactions of customers, a recurrent mining context for association rule analysis algorithms. The initial formulation of the model uses the "support-confidence framework", which is considered the "plain" formulation of rule miners. As time has passed, multiple proposals for improving the initial formulation have been done, after noticing the approaches might evolve into a powerful ML technique.

Although far away in terms of usage to other unsupervised learning algorithms such as clustering and dimensionality reduction techniques, association rule miners provide a very different approach that can be applied to datasets historically difficult to mine. Over the years they have been rapidly adopted to mine rules in a broad range of datasets and fields.

Recently and due to the importance of the field, large amounts of health data are being recorded, devoted to understanding those conditions in our body that are not seen at a glance. Medical data is one of the aforementioned datasets difficult to mine, comprising large amounts of data related to health, life, and many other domains, also presenting interactions with other variables that might or might not be present in the dataset. The acquaintance of the outcomes a person can experience based on their prior clinical issues is crucial not only for the early detection but also for medical resource-saving. Related to this, morbidity detection is a key concept to detect this interaction between conditions and outcomes.

A *trajectory* in the medical field, relates to the sequence of events/diagnoses/conditions that occur within the life of a patient. Some of these diseases are chronic and are demonstrated to present some correlation between them, co-occurring in some cases in a controlled period of time. It is crucial for the medical community to detect these co-occurrences. By collapsing all the trajectory data in medical databases, one could recreate a map of associations of diagnoses to guide medical knowledge in individual cases, always relying on probabilities.

Since the raw data in these cases are usually medical codes, stored in large strings, those datasets present a particular shape usually difficult to mine for standard unsupervised learning algorithms. Association rules

could be a good approach not only to solve the issue, yet providing interesting information from the medical point of view.

It must be emphasized that the objective of rule miners is to discover associations that might not be known prior to the analysis. This unsupervised nature makes them prone to possible errors and difficult to validate the results when lacking expertise of the field. The validity of the models must be done from a subjective/medical point of view, assessing if the rules retrieved are feasible.

Rule mining has previously been used in the context of "health data mining". Some articles propose the approach to identify patterns in apnea events in [Pombo et al., 2017] or liver disease in [Kumar and Sahoo 2013], focusing on a specific disease context. In [Lakshmi and Vadivu 2017], association rules are used to mine large volumes of "Electronic Health Records" (EHRs) to find correlations among diseases, symptoms and drugs. Lift is used to discriminate among interesting rules. In [Doddi et al., 2001] the interestingness is in finding the relationships between procedures and reported diagnoses in EHR. Association rules are used to extract knowledge in diabetic data repositories in [Stilou et al., 2001]. In [Brossette et al., 1998] association rules are used to extract interesting patterns in hospital infection control and public health surveillance, focusing on the case of infection of *Pseudomonas aeruginosa*.

Focusing on the case of morbidity research, governed by the large volume and sparsity of data needed to evaluate, some studies have shown promising results. Likewise, some studies are focused on finding comorbidity related to a target disease or group of individuals. In [Held et al., 2016] seventeen already known comorbidities were analyzed with several algorithms measuring interestingness according to an index disease in elderly men. In [Thai et al., 2009], a comorbidity study of ADHD based on rule mining has been undergone in the National Health Insurance Database of Taiwan. Same happens in [Wang CH et al., 2019], where the target diagnoses to study the morbidities are mental disorders.

Other cases intend to discover morbidities and multi-morbidities in global EHR databases, not focusing on target diagnoses nor groups. In [Lakshmi and Vadivu, 2019] a novel approach based on weighted association rule mining is used to discover comorbidity patterns.

1.2 Proposal

Despite the potential of the rule mining methods, the broad range of approaches tend to suffer from common issues. First, the dimension of candidate rules to explore grows exponentially with the growing universe of possible itemsets, making the execution time grow proportionally. The second issue is related to the output: it is difficult to achieve a human-readable amount of rules while don't fall into easy obvious ones, we will define this issue as the *interpretation-truism tradeoff issue*. Therefore, it is needed some expertise and even luck to set the parameters correctly to find those rules governing the data that are not known previously by the user. In addition, this parametrization may change from one to another datasets.

Over the years, several proposals have intended to address these hardships. The first issue was tackled by introducing a support threshold. The solution is based on the acceptance that not all candidates can be considered within reasonable running times. These allowed designing efficient frequent set miners and rule mining algorithms by exploring only those item-sets appearing "often enough". Therefore, only those whose relative frequency among the set of transactions exceed a certain ratio are considered as possible target rules, reducing significantly the universe of possibilities and therefore the execution times. Although efficient, this solution puts on the shoulders of the user the responsibility to choose the correct parameters to get the best possible rules, worsening the second issue.

The second issue is harder to tackle than the first one. In the standard *"support-confidence framework"* it is easy to check using any dataset or association miners, that high demanding thresholds yield few obvious rules ,while softening them leads to large amounts of redundant rules that cannot be interpreted by a human. Multiple proposals from different researchers, some of them still being formulated and with no consensus for the right one were proposed. Most of these methods are based on the notion of *"redundancy"* among rules, trying to reduce the set of rules to an non-redundant base of absolutely minimum size, the *Representative Rules*. Nonetheless, even when taking into account redundancy the results tend to be unsatisfactory for most of the proposals.

By pushing harder the intuition of redundancy and reaching a definition for *novelty*, better results are achieved. Some examples of algorithms proposed as measures of novelty through the extent that confidence of a rule is "robust" relative to the confidence of other related redundant rules are *confidence width* or *rule blocking* [Balcázar 2009]. It is empirically demonstrated that better results are obtained when using these approaches.

In this paper, we will use yet another algorithm formalizing the notion of "novelty" to detect rules, formulated in [Balcázar 2013], the confidence-boost. The algorithm encompasses at once both the bound over the *confidence width* and the ability to detect if a rule will be blocked efficiently. There is a plain approximation of the model and another one taking into account the closure space implicit in the data, with the aim to provide efficiency to the plain formulation. These approximations to rule miners still lack the parameter automatization, still requiring the tuning of the parameters based on the expertise or knowledge of the data by the end-user. Nevertheless, in [Balcázar 2013] a tool called yacaree, implements in python the confidence boost algorithm self-adjusting the support and confidence boost, thus improving the issue.

We will empirically demonstrate and test the improvement when using confidence boost in a real case dataset of medical EHR trying to mine associations between diagnoses. Evaluating the reducing and the validity of the rules provided by the tested algorithm. The dataset contains three and a half million visits from a reference population of 400.000 patients in the Catalunya-Central Medical Area. From the visits and the diagnosing, a medical history of each patient is constructed to extend the notion of *medical trajectory*. The trajectories can be grouped based on factors such as sex or age to retrieve rules representing each population segment.

A comparison of the results achieved with the algorithm and another classical rule miner will be done. For validation purposes we will compare the algorithms based on the execution times, the number of rules find and by using medical expertise to check the validity of the rules.

A final proposal is given to suggest a method to detect directionality between diagnoses within a rule, to detect temporality and order of occurrence. Besides, hierarchical clustering is used to extend the notion of morbidity and all its related terms “patient complexity” and ”multimorbidity”.

1.3 Objectives

The main objective of this thesis is to provide an unsupervised learning technique that could be useful to tackle several research questions in the medical field. Formally, the goal is to provide a method able to evaluate high volumes of medical data and gain knowledge, providing a “static photography” summarizing the associations between medical issues that a certain segment of the population undergoes, getting rid of all those conditions suffered by chance. This “photo” must be able to raise the knowledge from an individual point of view to a population point of view, giving summarized parameters that can be useful to predict possible outcomes for a certain individual.

By using a new propositions of association miners the objective is to detect which associations of diagnoses experience the population. These associations could be related to morbidities, ideally bringing to light comorbidity between diagnoses relating conditions and chronic diseases that might be object of analysis by the medical community.

The ultimate goal of the results is that them should act as a guide for the medical practitioner to detect which diseases or interactions between them have a higher prevalence in the medical system at a time or for a certain patient with a condition, in order to shift the decision-making in terms of resources and effort towards them. Moreover, the output should highlight possible unknown or unclear interactions between diagnoses that would be not detected ”with the naked eye”, providing new medical researching paths. Moreover, the output should be easy to interpret for practitioners, using techniques

The objectives tackled in the study are:

1. Evaluate the validity of the absolute boost formalization with respect to other rule miners.
2. Validate the correctness of the results in the context of the medical database being used in the study.
3. Evaluate the validity of the technique as a method to provide a whole image of the medical issues experiencing the whole population.
4. Evaluate the validity of the technique as a method to provide a whole image of the medical issues experiencing different groups of the population (based on different confounding factors).
5. Evaluate the validity of the technique as a method to extrapolate morbidity in patients.
6. Report the possible drawbacks experiencing the model to refine it.
7. Validate the usability of the technique from the medical perspective.
8. *Explore the effectiveness of the technique in the COVID-19 outbreak dataset.*

Chapter 2

Association rules in a nutshell

The goal of this chapter is to clarify the minimum knowledge needed to understand association rule mining algorithms and the proposed confidence boost algorithm from a mathematical point of view. The chapters cover the basic and initial intuitive formulations of rules and rule mining algorithms found in [Agrawal et al. 1996] and underline the general features of the topic of association rule mining. This chapter provides the reader with the basics of the *a priori* algorithm, the most intuitive version of the association rule mining approaches, that will be tested against the approach used in this thesis.

2.1 Definitions

To conduct all the definitions in association rule mining we will define the mathematical formulation's usage. We will define the set of available items denoted by U. Its subsets are called itemsets, denoted using capital letters from the end of the alphabet (X, Y, Z).

$$X \subset Y \text{ denotes proper subset while } X \subseteq Y \text{ denotes improper subset.}$$

To put in context the methodology for a better understanding, we will assume the existence of a finite well-defined dataset D consisting of n transactions, each of which is a subset of U.

Rules are defined following the notation $X \to Y$, meaning transactions in D having X *"tend to contain"* Y. For simplicity, we will denote *LHS* to the left-hand side of the rule, with the same notation in the case of *RHS*, the right-hand side. When talking about single items, and not itemsets, we will use capital letters from the start of the alphabet $(A, B, C, ...)$.

Support of an itemset will be denoted by $s^D(X)$ while the confidence of a rule will be denoted by $c^D(X \to Y)$. From now on, we will assume the dataset D is known and clear from the context. Therefore, there is no need to supperscore it in support and confidence. γ and τ will denote support and confidence thresholds respectively for ease of use.

For a better mathematical understanding, when writing rules we will make explicit always what part of the consequent is already in the antecedent and write all our association rules as $X \to XY$ where $X \cap Y = \emptyset$. In the implementations and in the outputs, at the time of showing the rule only the Y part is shown, consequently outputting the format: $X \to Y$.

2.2 Initial formulation

Association miners search for valid expressions of the form $X \to Y$, meaning the transactions containing X *"tend to contain"* Y, widely known as association rules. These expressions are defined by means of different parameters; in the "plain" initial approach:

- The $support(X)$ of itemset X is the directionality of the set of transactions containing X (the number of transactions containing subset X). An alternative less used is the normalized support $(X)/n$.
- The confidence of a rule is denoted by $c(X \to Y) = s(XY)/s(X)$, an empirical approximation to the conditional probability (how frequent is Y among all transactions containing X). Confidence is a natural selection when we want to prune and rank the output of the association rule mining algorithm, but can also be done employing support, lift...

Other approaches also take into account the lift. To cope with accepted reasonable running times, a pruning is made via thresholds to the initial space of possible rules to consider, only keeping the interesting ones.

- In many cases, we assume the context imposes a threshold on the confidence constraint $c(X \to Y) \geq \gamma$ over the rules, and likewise a support threshold constraint $s(X \to Y) \geq \tau$. Full-confidence implications are retrieved via non-strict inequalities with $\gamma = 1$, with the form $\emptyset \to Y$.

Definition 1 *When the confidence of a rule reaches 1, we call it an implication (all transactions containing X also contain Y). When it falls below 1, it is a partial rule.*

Proposition 1 *In this preliminary approach to the model, we will allow the antecedent subset to have the form $X = \emptyset$. In this particular case, the confidence will be equal to the normalized support $c(\emptyset \to Y) = s(Y)/s(\emptyset) = s(Y)/n$.*

In the initial approach for the rule mining algorithm in [Agrawal et al., 1993] association rules are restricted to $|Y| = 1$. Using this definition, the algorithm only considers rules with 1 item in the RHS, directly provided from each frequent set, therefore reducing the space of possibilities and boosting up the execution times. Nonetheless, using this approach will deprecate the results:

Definition 2 *Confidence 1 implications assume $A \to B$ and $A \to C$ are equivalent to $A \to BC$. Nevertheless, it is much more informative the later case, stating that B and C appear jointly often with A.*

The real format of the rules that we will use from now on is different from the mathematical point of view with respect to the ones that most of the algorithms shows as outputs:

Proposition 2 *Through confidence and support, rules $X \to Y$ and $X \to XY$ are equivalent in almost all the statements (as some part of the LHS is repeated in the RHS). When considering lift, this assumption is not matched.*

Definition 3 *The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. Some studies define the lift as a variable underlying the importance of association between subsets. The formula corresponds to $Lift(X \to Y) = \frac{s(XY)}{s(X)s(Y)}$*

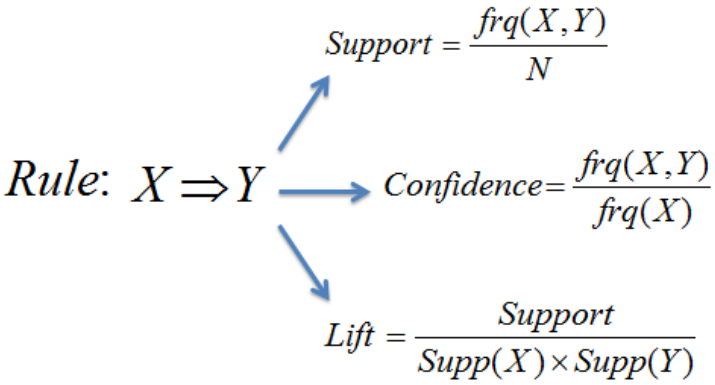

Figure 2.1: **Association rules parameters.** The figure represents the formulation of the parameters defining the primal formulation of association rules in [Agrawal et al,. 1993].

Several parameters can act as pruners for rules. By imposing thresholds on these parameters we control the number of rules that pass these limitations and are provided in the output. Figure 2.1 shows the principal parameters used in the plain formulation of the association rule miners.

Example 1 **At this point, devoted to an easy interpretation of the definitions and examples we will employ as running examples through most of this paper the closure space obtained from a specific dataset. For this example, the universe U includes the five items A, B, C, D, and E. The dataset consists of 12 transactions, six of which include all of U; two more consists of ABC, again two transactions consist of AB, and then one transaction consists of CDE and another one consists of BC.*

2.3 Limitations

With the initial formulation of the association rule mining task, it arises a twofold limitation.

First, the quantity of candidate itemsets of association rules $X \rightarrow Y$ grows exponentially with the often already large universe of items. The problem is related to the quantity of candidate itemsets to form the rules, which is essentially a combinatorial problem. It was solved by introducing a support threshold parameter. There, exploration is limited to those itemsets that appear "often enough" as a subset of the transactions. However, these solutions put into the shoulders of the user the expertise of choosing the support threshold, with little or no guidance. In [Agrawal et al,. 1993] the method imposes a threshold over the confidence (threshold on the conditional probability of Y conditioned to X).

Indeed, association rule mining aims to enumerate all the rules that are not disproved by the data. This standard approach, the "support and confidence" framework, it is well known tos suffer from the following issue: whereas high demanding thresholds yield few somewhat obvious rules, when softening them would lead to numerous redundant rules. Therefore, the need to have expertise and the luck of the user is crucial in this cases. A new proposition of the algorithms should be able to remove from the shoulders of the end-user the commitment of choosing all these parameters thresholds.

The second problem is related with output dimensionality. Often the set of rules provided as output is too large, especially if considering that its purpose is to be read and understood by the human. This problem has not received adequate attention. In the following chapter, a new proposition of rule miner algorithm attempts to provide yet one more approach to it, trying to solve this issue.

Chapter 3

Confidence Boost

The following chapters aim to introduce the mathematical notions used in the rule mining algorithm tested in the following parts of this thesis. First, an initial intuition of redundancy and novelty acts as a basis for the proposal used. These notions are the basis from the initial formulations extending the notions of novelty from [Balcázar 2009] to the final used formulation of confidence boost [Balcázar 2013]. The explanation uses the knowledge from *confidence width, blocking threshold and support ratio*, first propositions of novelty detection, to the *confidence boost* proposition used in the following chapters.

The last chapter introduces the framework designed for the ease of usability of the proposed algorithm, the *yacaree* program form [Balcázar 2015].

3.1 Novelty vs. redundancy

In association rule mining, novelty is approximated through a variant of the intuitive idea of redundancy. Several notions for redundancy exists in the rule mining field. In [Luxenburger 1991] redundant rules are those whose confidence can be computed from the confidence of other rules. Similar notions are studied in [Zaki 2004], but the approximate bases are constructed as rules having minimal generators at both LHS and RHS.

For the formulations of rule miners that we will use in the evaluation section, considering that the set of frequent closures is kept so that confidences are easily computed from them, we can reach a very simple yet precise notion of redundancy:

Definition 4 *Considering two rules, $X \to XY$ and $X' \to X'Y'$, the following are equivalent:*

1. $c^D(X' \to X'Y') \geq c^D(X \to XY)$ *and* $s^D(X' \to X'Y') \geq s^D(X \to XY)$ *for every dataset D.*
2. $X' \subseteq X \subseteq X'Y' \subseteq XY$.

When these cases hold, $X \to XY$ makes $X' \to X'Y'$ redundant so that the first is logically stronger than the latter. Every rule that reaches the confidence threshold and is not made redundant by any other rule above the threshold is part of the representative basis for that threshold. Hence, a rule is redundant because we can know beforehand, from the information on the representative basis, that its confidence will be above the threshold. Statement (2) implies statement (1). Accordingly, to [Balcázar 2010c]; the converse implication proofs that the representative basis has the minimum possible size among all bases for this notion of redundancy.

Proposition 3 *It is important to note that a representative rule at a given threshold may not be so at lower confidence, where some other rule might make it redundant.*

Proposition 3 is the basis to understand the next confidence width algorithm.

3.2 Confidence width

Assuming that the dataset D is clear from the context and a support threshold τ has been fixed, all our rules on the representative basis must reach at least τ on D. However, it is often the case that the representative basis provides several rules, making the human inspection unfeasible. Confidence width is a natural approach to cope with the redundancy in the representative bases.

As stated in [Balcázar 2009]; "A non-redundant rule of confidence c belongs to the basis for confidence threshold $\gamma = c$ if no rule of that confidence or higher makes it redundant. Equivalently, all rules that make it redundant have lower confidence."

Definition 5 *For an association rule $X \to XY$, consider all rules redundant to it, and pick the one with maximum confidence in D among them, say $X' \to X'Y'$. The confidence width of $X \to XY$ in D is:*

$$w(X \to XY) = \frac{c(X \to XY)}{c(X' \to XY)}$$

From definition 5 is easy to see that the condition that $X \to XY$ is redundant concerning $X' \to XY$ implies $c(X' \to X'Y') \leq c(X \to XY)$, hence the confidence width is always 1 or larger. $w(X \to XY)$ is slightly higher than 1 if and only if $X \to XY$ is a representative rule.

Therefore, for a confidence threshold γ and a rule $X \to XY$ of confidence c_0 the possibilities are:

- If no stronger rule appears at threshold $\gamma = c_0$, then $X \to XY$ will belong to the representative basis for that threshold.
- If a logical stronger rule appears at threshold γ, saying with a confidence c_1, with it very close to c_0, then the rule $X \to XY$ is not very novel, thus its confidence width will be barely above 1.
- On the contrary, a stronger rule may take longer to appear. In this case, only rules of much lower confidence entail $X \to XY$, so that the fact that it does not reach c_0 is novel in this sense. Making the confidence with higher above 1.

Subsumming it all in a single formula:

$$w(X \to XY) = -\frac{c(X \to XY)}{max\left\{c(X \to XY) | (X \to XY) \neq (X' \to X'Y'), X' \subseteq X, XY \subseteq X'Y'\right\}}$$

assuming again $X \cap Y = \emptyset$ and $X' \cap Y' = \emptyset$.

For fixed support, there may be rules that are not redundant to any other (all candidates may have lower support than that established by the threshold). By convention, we use ∞ as the value of the confidence width in those cases.

Definition 6 *The value of $w(X \to XY)$ is finite and well-defined if and only if either $X \neq \emptyset$ or Y have some proper superset Z with $s(Z) > \tau$.*

In [Balcázar 2009], some intuitions suggest that for a confidence threshold γ a natural choice will be setting the confidence width threshold at $w = 2 - \gamma$. However, the formal support to select is not very clear, and may change depending on the dataset.

3.3 Blocking rules

First proposed in [Balcázar 2009] and based in many prior propositions in [Bayardo et al., 1999; Liu et al., 1999; Padmanabhan and Tuzhilin 2000; Shah et al., 1999; Toivonen et al., 1995] referencing the notion that a subset of the antecedent may *"block"* an association rule. That is if the confidence of the rule with the smaller antecedent and the same consequent is higher enough.

Considering the association rule $X \to XY$, and reducing the antecedent to a smaller $Z \subset X$. Due to human intuition -with the natural habit to work with full implications-, one could think that the rule with the larger antecedent should be subsumed by the other, but this is not necessarily mate with association rules.

For instance, it is easy to check that at confidence 1, if $AB \to C$ holds, $A \to C$ also holds, not bringing new information. Association rules aren't implications, they relate relative frequencies. Indeed: rule $X \to XY$ speaks about the abundance of Y among the population of transactions containing X. Reducing the antecedent into Z, changes the population into, in principle, a larger one, and Y can be distributed at very different rates along each of these transactions.

Example 2 *Consider rules $A \to C$ and $AB \to C$. Hypothesizing a situation where almost all transactions with A and B have C, but there is a small fraction of them having A, thus the confidence of $A \to C$ is very small whereas that of $AB \to C$ is almost 1.*

As a consequence, we can state the fundamental that confidence does not detect negative correlations.

Example 3 *Fixing a confidence threshold of* 0.75 *and considering a D of n=10: 3 BC, 6 C, and 1 B. Then $c(B \to BC) = 0.75$ reaching confidence threshold. Most miners would report $B \to C$ as interesting in that threshold. However, the correlation between B and C is negative. Indeed, C is less frequent among the transactions having B than in total population: $c(\emptyset \to C) = s(C)/n = 0.9$.*

To solve this situation, the natural reaction is to divide the confidence by the (normalized) support of the consequent of the rule, eventually getting the lift (also named as *interest* or *strength* in the literature). The methodology is a fundamental from probability, as measures the deviation from independence, as a multiplicative distance from the case of fully independent X and Y which would give value 1 for it:

Definition 7 *The lift of rule $X \to Y$ is:*

$$lift(X \to Y) = \frac{c(X \to Y)}{s(Y)/n} = \frac{s(XY)n}{s(X)s(Y)}$$

It must be noted that contrary to confidence, the lift of $X \to Y$ does not coincide with that of $X \to XY$. If we are to use the lift, then we must be careful to keep the right-hand side disjoint from the left-hand side: $X \cap Y = \emptyset$. Note also that, in the case of $X = \emptyset$, lift trivializes to 1.

Although lift is an interesting measure, lacks the ability to orient rules, since both sides are symmetric. Additionally, is limited to control cases where $c(Z \to Y) > c(X \to Y)$ for $\emptyset \neq Z \subset X$.

Example 4 *In [Balcázar 2009] it is pointed that in the ADULT dataset from Irvine [Asuncion and Newman 2007] at 5% support and 100% confidence 67 out of 71 rules are of the form "Husband" + something else → Male, with the others being the same with another antecedent. Rule "Husband" → "Male" is not found since one of the tuples has "Husband, Female" therefore deprecating the rule at confidence 100%. This opens the door to many other rules, that would be blocked, enlarging the LHS. The whole issue will not be solved by dividing the confidences by the support of "Male".*

From example 3 it is clear the needing to react to negative correlation problem for confidence while still maintaining guidance. In [Balcázar 2009] it is proposed a *blocking threshold* using the confidence at which a smaller antecedent would block the rule.

Definition 8 *Giving rule $X \to XY$, $X \cap Y = \emptyset$, proper subset $Z \subset X$ blocks the rule at blocking threshold b if:*

$$\frac{s(XY) - c(Z \to ZY)s(X)}{c(Z \to ZY)s(X)} \leq b$$

The paper suggests b values to be positive around $\leq$ 0.2.

The intuition in 8 is based on discarding rule $X \to XY$ in case finding a rule $Z \to ZY$ with $Z \cap X$, having almost the same confidence, or larger (in the presence of support threshold, both rules must surpass it). Therefore, if the larger rule does not bring high enough confidence concerning the simpler one: it remains blocked.

The higher the blocking threshold is, the more demanding it becomes the constraint to rules. From the paper, it is pointed that the problem of the ADULT dataset aforementioned is solved by using a very small threshold ($b = 0.000075$). A proposal for parametrization is, being b the confidence width bound, then a probable correct threshold is around $b - 1$.

Example 5 *$\emptyset$ blocking $A \to BC$ in example 1:*

$$\frac{s(ABC) - c(\emptyset \to BC)s(A)}{c(\emptyset \to BC)s(A)} = \frac{8 - (9/12)10}{(9/12)10} = 0,066$$

3.4 Support ratio

The notion of suppport ratio is related with the parameters of confidence width, rule blocking, and absolute boost. First employed in [Kryszkiewicz 2001], it is intended to provide a faster algorithm to compute representative rules. As demonstrated in [Balcázar and Tirnauca 2011], the approach is efficient and useful but may run into the risk of providing incomplete output.

Definition 9 *In presence of support threshold τ, the support ratio of rule $X \to XY$ is:*

$$\sigma(X \to XY) = \frac{s(XY)}{max\,\{s(Z) | XY \subset Z, s(Z) > \tau\}}$$

As before, by convention, the value will be set to ∞ if $Z = \emptyset$.

Proposition 4 *If the value of $\sigma(X \to XY)$ is finite and well-defined then $w(X \to XY)$ is also finite and:*

$$w(X \to XY) \leq \sigma(X \to XY)$$

It can be noted that $\sigma(X \to XY) \geq 1$ for all rules; being 1 for XY if it is not closed, since these sets are those having a proper superset Z with the same support. To the practice, the formalization of the support threshold is important since many of the quantities we study for an association rule are bounded from above by the known support ratio, and, therefore, will trivialize to value less than or equal to 1 unless we consider only closed sets XY as RHS.

3.5 Confidence boost

Definition 10 *Confidence boost of rule $X \to XY$ assuming $X \cap Y = \emptyset$ is*

$$\beta(X \to XY) = \frac{c(X \to XY)}{max\,\{c(X \to XY) | (X \to XY) \neq (X' \to X'Y'), X' \subseteq X, Y \subseteq Y'\}}$$

As in previous definitions, rules in the denominator are required to surpass the support threshold. Again by convention, if subset in the denominator is empty confidence boost reaches infinite.

Proposition 5 *The value of $\beta(X \to XY)$ is finite and well-defined if and only if either $X \neq \emptyset$, or Y has some proper superset Z with $s(Z) > \tau$. The set of rules with infinite confidence boost is the same set of rules with infinite confidence width.*

Example 6 *Considering example 1, rule $A \to AB$ has a confidence boost of 16/15. Considering all rules $X' \to X'Y'$ with $X' \subseteq A$ and $BC \subseteq Y'$; one can see that the maximum confidence among them is 0.75, attained by $\emptyset \to BC$. Then $\beta(A \to BC) = 0.8/0.75 = 16/15$.*

Similar to the notion of confidence width, a low confidence boost tends to low novelty. Explained as follows: for a confidence boost value b, slightly higher than 1; according to the definition, the confidence of $X \to XY$ is not much higher than that of $X' \to X'Y'$. All transactions having X do have X' and all transactions having Y' do have Y, therefore the confidence of $X \to XY$ is not that novel, not giving much additional confidence over a rule that states a similarly confident, intuitively stronger fact $X' \to X'Y'$.

As an orientation, we should not consider rules with a confidence boost equal to or lower than 1. This solves the objection against confidence that negative correlations remain undetected: if support of B is 0.8 a rule $A \to B$ of confidence less than that would yield a confidence boost below 1, due to rule $\emptyset \to B$.

Although the representative rules are stated as a minimum size basis without redundancy, often comprise many rules. It must be assumed that when using thresholds as well as when using the confidence boost we are losing information in the ease of interpretability. Most of the papers [Luxemburger 1991; Pasquier et al. 2005; Zaki 2004; Balcázar 2010c] treat different implications, which allow for more compact bases, from the partial rules. In [Balcázar 2013] the method uses a notion of closure-based redundancy which provides a complete basis of provably minimum size denoted by $B*$.

The main advantage of using this definition is that it provides bases of a size comparable to the representative basis $R*$. This turn to have the property to compute faster these part of the $B*$ basis referring to the partial rules (with confidence below 1). Moreover, the $B*$ basis can be computed from only the closures. Nevertheless, this boosting will never compete with the velocity of computation of those miners using the standard *confidence-support framework*.

3.6 Yacaree

Yacaree is the open-source tool implementing the closure-based confidence boost rule miner. The goal is to create a tool that can be used by anyone with little to no guidance about the parameters to set up. Yacaree *(Yet Another Closure-based Association Rule Experimentation Environment)* is fully implemented in python. The key properties are the self-tuning of support and confidence boost thresholds when mining high-boost $B*$ association rules.

The method pursues an initially very low support bound, and progressively increases it. Frequent closures are mined via a variant of the *ChARM* method [Zaki and Hsiao 2005], similar to a depth-first search but ordering closed item-sets in decreasing support, increasing the performance when support threshold is higher while not invalidating rules found so far.

The idea is similar to the one found in *a priori* approximation [Witten and Frank 2005], a method that will also be tested in the following parts, but with the need to set up some parameters such as the min-max bounds over the support threshold. Moreover, in a priori algorithm the desired number of rules to get by the user is set before the run, presumably ignoring interesting ones, and not finding rules of low support. In the case of yacaree, the user can decide to set a desired number of rules or explore all the closure-space to found all associations surpassing the thresholds.

Each closure found in each closed set is analyzed, searching for alternative extensions without failing the current support threshold. From the extensions, the one providing largest support-closed set is retrieved.

As pointed out in [Balcázar 2013], the closures are translated to a lattice construct based on the procedure in [Baixeries et al. 2009] used to make available immediate predecessors of each closed set for computing the basis $B*$. Rules are constructed from the lattice and discarded if it is guaranteed that a future threshold adjustment will never recover them, processed if obey thresholds or maintained if may obey threshold after future adjustments.

Support threshold starts at a trivial level until reaching memory consumption dependencies (when the heap where unexpanded closures are stored its considered in overflow). At that point, minimal support constraint are raised until computation is able to continue.

From the confidence boost point of view, a standard non-demanding confidence threshold is set so it lets large quantities of rules to pass the constraint, that would be posteriorly pruned via a threshold on the closure-based confidence boost. The self-adjustment is made through the lift.

For the constructed lattice of closures, a certain closure is not adequate to yield high boost rules if its support ratio is lower than the current confidence boost threshold. Despite that, the target could become adequate to yield high boost rules if in the future the confidence boost threshold decreases. Therefore, the confidence boost constraint is partially "pushed into" the mining process by temporarily omitting the expansion of such closed sets.

Consequently, the target closures are maintained in a separated data structure from where they are "captured" in case a decrease of the boost bound promotes them to be candidate closures for creating high boosting rules.

The default mining process rely on an initial maximum confidence boost and a minimum threshold. The program starts with a demanding confidence-boost constraint. A rule must have 15% more confidence than any other rule on confidence boost in order to quantify it as interesting. This threshold can be tuned by the user.

In many datasets, this confidence bound is too demanding. In those cases, the program monitors the lift of rules with a single item as antecedent from a closed set with support ratio higher than confidence boost bound. If lift values decrease, they enter weighted average with current confidence boost, eventually decreasing it. By means of it, a tracking of the degree of correlation is found to reduce the confidence boost progressively, until reaching the lower bound 1.05 (meaning rules presented must be 5% more confidence than any other stronger rule).

Chapter 4

Data

This chapter is intended to contextualize the dataset used in the experimentation methodologies being held in this thesis. The chapter is also used to outline the preprocessing endured by the dataset. The dataset used was initially conceived to match the requirements of the medical community to identify how the diagnosing strategy has changed pre and post covid, with the initial objective to identify which conditions have suffered an underdiagnosing. Therefore, the dataset mined was devoted to another analysis different from the one depicted in this thesis. It is obtained thanks to a collaboration between *ICS-Catalunya Central* and *ICS-Metropolitana Nord*, which has lead to the published articles [Lopez Seguí F et al., 2021] and [Pifarré i Arolas H et al., 2021].

4.1 Dataset

The Public Catalan healthcare System provides universal coverage to 7.6 million inhabitants with an important role in community and primary health care. This analysis uses a database from the Primary Care Services Information Technologies System of the Health region of Central Catalonia (Catalonia, Spain) belonging to the Catalan Institute of Health. The objective of the study is a total of 3,555,799 visits from primary care corresponding to the years 2019 and 2020 (period covering most of the COVID-19 pandemic). The database comprises both face-to-face visits (at primary-care health systems or home) and teleconsultations (telephone and mail or via the app from ICS *la meva salut*) corresponding to 376,486 citizens (out of a total of 404,245 that make up the reference population of this health region). Regarding ethic considerations, the visits are anonimized matching the medical databases requirements.

Each visit is constructed by both "active diagnoses" (those diagnoses that are active at the moment of the visit, therefore mainly chronic diseases or diagnoses that are not solved by a single visit) and "visit diagnoses" (diagnoses of the visit). Based on the results from [Lopez Seguí F et al., 2021]: only 79,43% of the visits have "visit diagnoses". The degree of coding is slightly higher in the face-to-face visits compared to the teleconsultation visits (81,97% vs. 75,23% respectively). The diagnose codes have been aggregated corresponding to the reference ICD-10 dictionary [ICD-10-CM, 2019]. The codification can be discretized using a hierarchical definition in different granularities. The smaller granularity corresponds to specific conditions or situations of the illness. This granularity has not been used in the study since the specificity leads to several codes presenting little to no prevalence among the patients, thus loosing information. The granularity used corresponds to the concept of "disease", "sociocultural condition" or "visit definition" and does not take into account the particular conditions or situations of the illness (1532 different codes, with the format *Malignant neoplasm of anus and anal canal, C21*).

Moreover, a visit contains extra supplementary variables defining the patient (age, gender), the location of the health Center, and the type of the visit. The id of the patient is unidentified with security purposes.

Medically speaking, the year 2020 must not be compared to other years in the healthcare systems. COVID-19 outbreak has caused a disruption in the primary care model. In 2020 the increase in telemedicine (+267%) does not compensate the decrease in face-to-face visiting (-47%), with an overall reduction in the total number of visits (-1.36%) reaching a record (-7.56%) if excluding COVID-related visits. The social group of people as well as their diagnoses can be different regarding the type of visiting, is important to account for this in the study. Therefore, the stress suffered by the system may have caused a disruption in the coding strategy.

The purpose of the datasets is to test the proposed YACAREE approach. As aforementioned, the year 2020 must not be treated equally to the year 2019. Covid-19 outbreak has had a disruptive effect in all the medical services around the world. While the COVID-19 has changed the nature of the vast majority of the visits, some patients have not been diagnosed due to the fear to go to the medical centers (conceiving them as infection focus) and the already tensioned system has not been able to keep diagnosing and medical procedures up-to-date, possibly infra-diagnosing some groups or diagnoses. Moreover, the nature of the visits has been completely swung, presenting the record telemedicine visits in the year 2020, possibly affecting the diagnosing procedures.

Women account for 55,23% of the visits, fairly unchanged with respect to 2019. The average age of the patients has been reduced in 0.92 years (53,54 and 52,89 years respectively in 2019 and 2020), mainly due to the effect of the change to telemedicine visiting.

For the ease of use of the age variable, it has been discretized based on medical advice as follows: Childhood (0-11 years) , Adolescence (12-17), Adulthood (18-59), elderly (60-∞).

As a result, year 2020 must not be treated as a reference year for validation purposes, but could be a good target to find possible associations that have not been seen at a glance, possibly pointing out interesting sources of investigation for the medical community. When performing empirical and subjective evaluation of the model in section we will take into account the year 2019, which probably acts better as a reference year and present implications that would be interesting but known for the medical evaluation.

4.2 Diagnosing

As aforementioned, the subject of study from which the rules have to be computed is the ***"medical trajectory"*** of patients, the concept relies on the medical trajectory definition from [Zamora M and Gavaldà R., 2017], extending some different notions based on the input dataset. From now on, we will be talking about trajectories instead of transactions in rule miners.

Definition 11 *A medical trajectory is composed by all the diagnoses that codifies the health issues befallen throughout the life of an individual.*

In an ideal world, a patient will go to the medical center each time some health issue occurs, and the medical expert would correctly diagnose these issues. Nevertheless, this sometimes doesn't occur due to the stress of the medical system, the poor conciousness of the coding by the medical expert o the error of diagnosing. Moreover, in order to achieve an accurate medical trajectory, one should collapse data of medical visits expanding long in time, turning to be difficult in terms of data processing capabilities and database dimension.

From our dataset, it is not possible to define the whole medical trajectory, since we only have information regarding two years of medical visits. We have to construct the medical trajectory from the set of visits each patient has done to the medical centers. Each of the visits contains active and visit diagnoses.

- Active diagnoses are defined as those diagnoses that could be chronic or expand a long period throughout a life of a patient, involving non-resolved health issues.
- Visit diagnoses are those that are diagnosed in the visit, therefore associated with the illness succeding at the time.

By adding active diagnoses, we are expanding longer through time, seizing more info than just the diagnoses from each visit. Moreover, it is important that by means of letting the active diagnoses to enter to the trajectory we are able to find morbidities between non-resolved health issues and visit diagnoses. Even so, the passive closed diagnoses are not found in the dataset, avoiding some information and possible morbidities. It is important to note that the active diagnoses are found in the dataset ordered by importance. Therefore, there is a part of the visits that does not follow a temporality, thus loosing the order of appearance. Ideally, a trajectory should be able to be ordered in time to improve the morbidity research.

Our notion of trajectory does not present repeated diagnoses. One limitation of association rule mining is that the presence of a duplicate item in a dataset does not have overall effect on the final output. The support of a frequent itemset is based on the number of rules presenting the itemset, not the number of total appearances of the itemset in the database. Therefore, it is not useful to have duplicated items in a trajectory. Nevertheless, in posterior steps we will need the dataset not to present repeated items. We will keep the first occurrence of these duplicated diagnoses, assuming most of the time, the repeating of one medical issue could be due to a non-closed condition.

The trajectory construction algorithm will be therefore:

Algorithm 1 Trajectory construction algorithm

Data: Dataset D; n users; m visits per user defined by date. Each visit contains:
$A = \{a_1, ..., a_k\} \neq$, Set of unique **active diagnoses**.
$B = \{b_1, ..., b_k\} \neq$, Set of unique **visit diagnoses**.
Result: A character vector containing the trajectory diagnoses separated by spaces.

```
for user in Users do
    for visit in Visits do
        if visit < ∃Visits then
            Trajectory = A_visit ∪ B_visit
        end
        else
            SubTrajectory = A_visit ∪ B_visit
            Trajectory = SubTrajectory ∪ (B \ A)
        end
    end
    return Trajectory
end
```

The diagnose codes are extracted from the international classification of diseases ICD-10. This classification presents a hierarchical format of diagnoses. The Catalan healthcare system information systems are also based in this classification. This hierarchy presents different granularities based on the specificity of the diagnoses, going from 21 chapters to over 90.000 codes in the smaller possible coding granularity. The ranking hierarchy is the summarized in 4.1:

C00-D49 - Chapter 2: Neoplasms	A
C60-C63 Malignant neoplasms of male genital organs	**B**
C60 Malignant neoplasm of penis	**C**
C60.9 Malignant neoplasm, non-specified location of penis	D

Table 4.1: Different granularities present in CIM10.

The raw dataset presents the codification at **D** level. Nevertheless, although this granularity is the one most informative it suffers in support computation, since most of the codes present small prevalence in the dataset. A proper trade off between the informativeness of the granularity and the number of codifications is mandatory to achieve a good performance in rule mining and a proper interpretation.

In the testing part, the algorithm will mine granularities **B** and **C**, with 262 and 1673 codes respectively. A preprocessing step is done to collapse all diagnoses from the initial granularity **D** to the desired ones. Predictably, as the collapsing reaches higher granularity, it is often found that a trajectory presents repeated codes (this is easily explained because diagnoses **C60** and **C63** will both collapse to diagnoses **C60-C63** if found in the same medical trajectory). The repeated items are removed from the trajectories as aforementioned, keeping the initial appearance.

It must be underscored that the trajectories could take into account not only active and visit diagnoses, yet other variables defining the patient, such as age or sex (named confounding factors). With this purpose, age is discretized in: *Childhood (0-11), Adolescence (12-18), Adulthood (19-59) and Elderly (60 – Inf).*

The raw trajectories will have the form:

sex=female age=adult A09 F17 F41 H61 I10 I11 I34 I35 K21 M54 R01 Z63

And the consequent anonymized trajectory:

A09 F17 F41 H61 I10 I11 I34 I35 K21 M54 R01 Z63

Concerning the trajectories accounting for the confounding factors, we will also create datasets representing a certain group of population based on sex and age. These datasets will have anonymized trajectories, since saving age or sex will lead with almost all rules presenting implications with respect to these variables. Therefore, the partitioned datasets are defined as anonimized trajectories of a single segment of population based on confounding factors.

Image 4.1 shows that the longer trajectories are those corresponding to the elderly population. This segment of population is the one that uses more health services and tends to suffer from a higher number of diseases, leading to longer and more informative trajectories. From childhood to old-age population the number of diagnoses grows accordingly to the age.

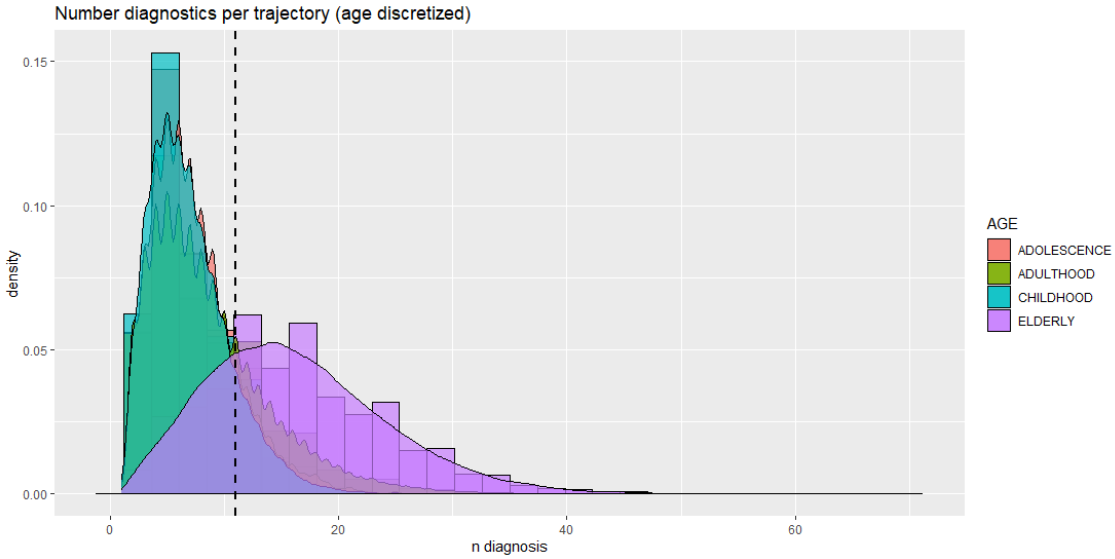

Figure 4.1: **Density plot trajectories length.** Density plot of the histogram defining the number of diagnoses per *medical trajectory* in the dataset, differentiated per age.

Based on this different propositions, with the goal to achieve more specificity and validity, we will test the following datasets:

- Dataset in granularity **B** with **RAW** trajectories **ALL** .
- Dataset in granularity **C** with **RAW** trajectories **ALL**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories **ALL**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **MALES**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **FEMALES**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **CHILDHOOD**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **ADOLESCENCE**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **ADULTHOOD**.
- Dataset in granularity **C** with **ANONIMIZED** trajectories of **ELDERLY**.

It is important to note that diagnoses codes are not only related to medical conditions, yet also can relate to socioeconomic/sociodemographic factors. One example are diagnoses *"Z55-Z65 Persons with potential health hazards related to socioeconomic and psychosocial circumstances"*. By means of these codes, we are not only creating associations between diagnoses, but also between patient complexity status. These codes are defined in table 4.2.

Code	Description (ICD-10)
Z55	Problems related to education and literacy
Z56	Problems related to employment and unemployment
Z57	Occupational exposure to risk factors
Z59	Problems related to housing and economic circumstances
Z60	Problems related to social environment
Z62	Problems related to upbringing
Z63	Others problems related to primary support group, including social circumstances
Z64	Problems related to certain psychosocial circumstances
Z65	Problems related to other psychosocial circumstances

Table 4.2: *Persons with potential health hazards related to socioeconomic and psychosocial circumstances* codes in ICD-10.

Chapter 5

Results

This chapter is intended to test the proposed association rule miner model in the aforementioned dataset, evaluating the correctness and possible issues arising from it. First, an empirical evaluation is done to check the performance of the confidence boost algorithm in terms of the output dimension. Second, a subjective evaluation tries to assess the correctness of the rules based on medical knowledge.

Different partitions of the dataset have been created based on the confounding factors of the patients (age and sex) to push harder the performance of the model, trying to improve the main drawbacks of it. First, an initial empirical evaluation will be held without undertaking any evaluation on the correctness of the rules, yet only studying the approach in terms of the amounts of rules presenting based on the thresholds.

For this evaluation procedure, we will only take into account the trajectories from 2019. Ideally, one should merge the datasets from different years to have a trustable reference year. Nevertheless, this is impossible in this case due to the large amounts of data and the difficulty to get data from public health sources. In the case of the subjective evaluation, we need to make sure the results in 2019 are trustable from a medical point of view to assure the procedure works in 2020. In the latter case, it might be possible to encounter unknown associations due to the pandemic.

It is important to highlight prior to deepen in the examinations of the results, that sometimes, the practitioner may not diagnose correctly. The deficit of medical practitioner in the Catalan health services lead to each doctor having a visit every 6 minutes, eventually reaching 12 in some visits. This, together with the poor awareness of the importance of the correct diagnoses coding in the Primary Care services (in contrast with the awareness in the cutting-edge hospitals of the Catalan health-care services) leads to miscoded diagnoses in some cases. This situation causes that some codes act as a hotchpotch, having high support because most practitioner over-code them, although may not be correct.

One example of the aforementioned is the code *"Z00 — Encounter for general examination without complaint, suspected or reported diagnoses"*. Several practitioner tend to over-use this code, in order to be able to meet the deadlines of the visits. Moreover, since they do not directly find examples where the coding provides an interesting feedback for them, they don't raise awareness about the importance of correctly coding.

All the rules retrieved by the algorithm have been evaluated by medical experts to prove the validity of the results from a medical point of view.

All the code used in the Methodology and Results phases of this thesis is available at: Github repository.

5.1 Empirical evaluation

In the paper presenting the yacaree [Balcázar 2013], the method is tested with the ADULT dataset (part of the Adult US census dataset from UCI [Asuncion and Newman 2017]), RETAIL dataset (from FIMI repository) and NOW dataset (based on Neogene of the Old World dataset [Fortelius 2003]).

As aforementioned, the datasets that we will use for the empirical evaluation of the model are those of medical trajectories in 2019 accounting for 195.661 patients, the 48,4% of the reference population in the area of study. Table 5.1 (based on the nature and format of the datasets used) gives interesting information about the dataset: The size (number of trajectories), the numbers of items involved in each, and the total number of item occurrences. Table 5.1 retrieves the information of the datasets used partitioned by the confounding factors sex and age.

	Parameters		
Dataset	Size (trajectories)	Items (diagnoses)	Occurrences
Raw (min granularity)	195661	1679	2502455
Anonimized (min granularity)	195661	1673	2111133
Anonimized (max granularity)	195661	262	2110334
Male	97372	1521	985241
Female	117648	1572	1412318
Childhood	17238	893	120909
Adolescence	12610	986	92247
Adulthood	113898	1589	1002610
Elderly	71274	1475	1181793

Table 5.1: **Test datasets Definition.** Non-partitioned and partitioned test datasets based on confounding factors sex and age definition: Test Datasets

It is important to highlight that most of the trajectories are from the adult population, due to the fact that comprise a wider range of ages comparing with other groups. The itemsets used have similar dimensions between childhood and adolescence ages and between elderly and adulthood ages, being the adulthood the segment presenting more different items.

Image 5.1 shows the most frequent items found in the overall dataset. Table 5.2 describes the top-10 conditions codified in it.

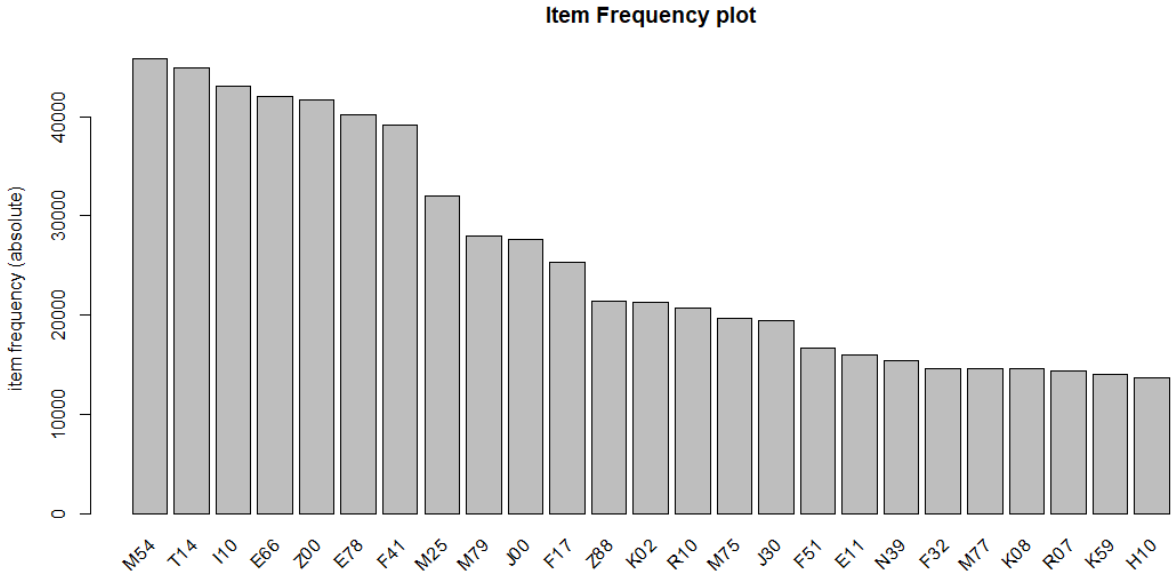

Figure 5.1: **Histogram of common diagnoses.** Most frequent diagnoses appearances in the *dataset with granularity C and raw trajectories.* Age, sex, yearly histograms

Code	Description (ICD-10)
M54	Dorsalgia
T14	Injury of unspecified body region
I10	Essential (primary) hypertension
E66	Overweight and obesity
Z00	Encounter for general examination without complaint, suspected or reported diagnoses
E78	Disorders of lipoprotein metabolism and other lipidemias
F41	Other anxiety disorders
M25	Other joint disorder, not elsewhere classified
M79	Other and unspecified soft tissue disorders, not elsewhere classified
J00	Acute nasopharyngitis [common cold]

Table 5.2: **Most frequent codes.** *Top-10 codes sorted by support in the initial non-partitioned dataset.*

5.1.1 Boost-confidence ratio

Each dataset has been mined at three different levels of confidence and 7 different levels of confidence boost. Tables 5.3, 5.5, 5.6 report, for each pair of confidences and confidence boosts, the number of rules retrieved by the model. The confidence boost thresholds act as pruners towards the rule basis found at confidence boost 1.00 for each confidence, reducing it to its minimum non-redundant expression passing the threshold. It is recalled that the support threshold is automatically set up by the data and the yacaree algorithm.

	70%			80%			90%		
β	RAW	ANO-min	ANO-max	RAW	ANO-min	ANO-max	RAW	ANO-min	ANO-max
1.00	7554	1599	1067	4253	661	343	2103	43	14
1.05	2143	672	417	924	214	56	288	29	0
1.10	1099	349	151	459	101	19	161	12	0
1.15	684	211	54	264	60	9	97	11	0
1.20	478	135	33	178	33	4	73	9	0
1.25	366	85	24	159	26	1	69	7	0
1.30	289	69	17	142	23	1	67	7	0

Table 5.3: Effect of confidence boost and confidence threshold on the output number of rules. *(RAW: RAW dataset ; ANO-min: ANONIMIZED min-granularity; ANO-max: ANONIMIZED max-granularity)*

It is obvious that both the confidence boost and the confidence threshold act as pruners reducing the number of rules when increasing them: the first by imposing a threshold to the novelty of the rule and the second to the confidence of the implication.

At a glance, from 5.3, it is important to note that when using the max-granularity, thus reducing the items in our dataset, the number of rules found drops rapidly. This is produced due to the fact that we are losing information with respect to the more fine granularity.

Table 5.4 shows the results of a rule at confidence 90% and absolute boost 1.30. First, it must be noted that when using the fine granularity the degree of the rules tends to be lower than in the coarse granularity. Nevertheless, although both rules seem to define correct associations from the medical point of view, the one for the fine granularity shows more accurate information. Rule 1 relates obesity, renal failure, and heart diseases with hypertensive diseases. On the other hand, rule 2 is pointing a direct relation between an under-dosing of some drug of the cardiovascular system with a hypertension episode.

Rule	**Parameters**	**Traduction**
{E65-E68, I30-I52, N17-N19} → {I10-I16}	[conf: 0.920; supp: 0.012; lift: 3.399; leverage: 0.917; PS: 0.009; S-S: 11.515; boost: 1.346]	{Obesity and other hyperalimentation , Other forms of heart disease, Renal failure} → {Hypertensive diseases}
{T46} → {I10}	[conf: 0.933; supp: 0.004; lift: 4.238; leverage: 0.932; PS: 0.003; S-S: 13.850; boost: 1.484]	Poisoning by, adverse effect of and underdosing of agents primarily affecting the cardiovascular system → Essential (primary) hypertension

Table 5.4: Evaluation specificity of rules in the two granularities used in the *RAW anonimized datasets*.

Based on this knowledge, the partitions of the dataset based on the confounding factors sex and age are held using the finer granularity. The resulting number of rules following the aforementioned methodology in the partition's dataset is subsumed in tables 5.5, 5.6:

	70%		80%		90%	
β	FEMALES	MALES	FEMALES	MALES	FEMALES	MALES
1.00	1825	1858	829	673	62	72
1.05	677	867	206	267	12	33
1.10	316	436	85	135	6	15
1.15	165	242	35	96	3	12
1.20	98	142	22	55	2	9
1.25	64	103	13	39	2	8
1.30	57	78	13	32	2	8

Table 5.5: Effect of confidence boost and confidence threshold on the output number of rules.

	70%				80%				90%			
β	A	B	C	D	A	B	C	D	A	B	C	D
1.00	5077	2444	301	3189	4623	874	144	1413	4144	150	45	91
1.05	1430	1941	209	652	1114	759	95	316	925	133	37	8
1.10	543	1281	144	132	308	593	63	44	159	108	25	3
1.15	351	755	102	75	171	427	49	16	57	89	15	3
1.20	276	514	86	62	128	267	45	15	42	60	13	3
1.25	218	208	71	44	92	146	39	12	34	42	12	3
1.30	167	114	62	27	59	77	35	12	22	30	12	3

Table 5.6: Effect of confidence boost and confidence threshold on the output number of rules. *(A: Childhood ; B: Adolescence; C: Adulthood; D: Elderly)*

Although having more trajectories from females (+17%) the amount of output rules is higher in males, possibly presenting clearer non-redundant associations.

At a glance, one can see that a higher number of rules are found in the case of adolescence and childhood, though having fewer transactions. We will see later that these groups such as elderly tend to be governed by a set of diagnoses since account for high support and the associations are clear. For the groups such as adolescence and childhood, one cannot found these clear implications, therefore the associations are not governed by a set of diagnoses.

It is important to understand that a higher dataset might not traduce in more associations. The associations are affected by the "clarity" of the dataset. In the case of the childhood segment, there are a lot of associations surpassing the thresholds, but the number is much lower in the case of elderly segment. This may indicate an increment on the threshods is needed in the case of childhood age. Moreover, it also indicates the trajectories in childhood age are not as clearly defined with respect to the ones on elderly population. By intuition, elder population tend to suffer from similar conditions and cooccurence of conditions induced by age, thus presenting itemsets of diagnoses with larger support and confidence. In the other hand, childs may enccounter the healthcare system by several unpredictible and unrelated conditions, caused by chance. Moreover, the effect of the trajectories lengths is also producing more unclear associations in this segment.

5.1.2 Algorithm evaluation

This subsection is intended to prove the performance of the proposed algorithm against other classical rule miner approaches in terms of the number of rules given based on the confidence and the confidence-boost thresholds. Therefore, evaluating the effect of the noverly and redundancy propositions present in yacaree. As the previous analysis, this section is not meant to underline the medical validity of the rules, but to empirically validate the procedure in terms of the output number of rules. The A priori algorithm has been carried using the minimum support used in yacaree.

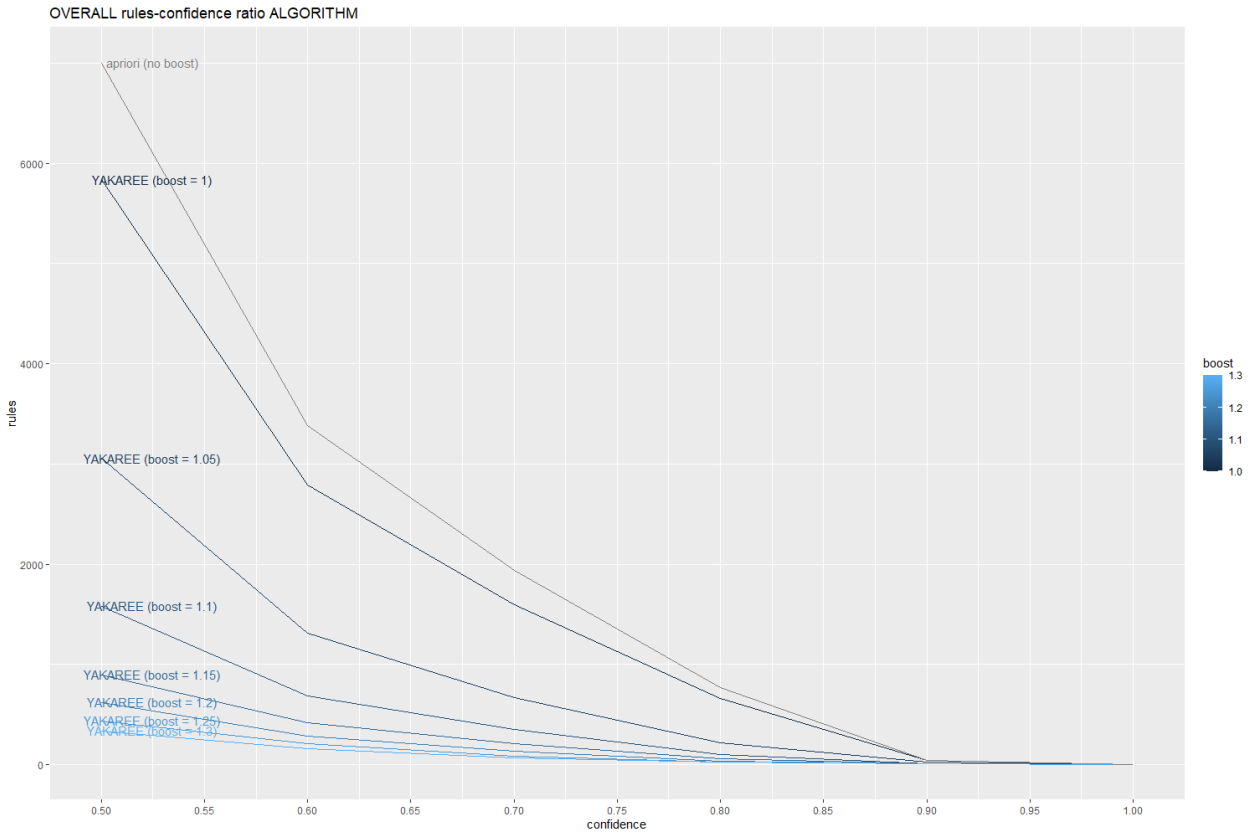

Figure 5.2: **Comparation Yacaree and Apriori rule miners.** Comparation of the two algorithms in the overall dataset for different values of confidence and boost thresholds.Outputs the number of rules

.

A major finding is that for confidence boost 1.0, the number of rules retrieved by yacaree is always lower than the number of rules retrieved by the “a priori” algorithm. The use of closure-based spaces reduces the possibility space to check on the yacaree, thus reducing the computation time. This improvement in terms of time can avoid the computations of some significant rules, therefore showing lesser rules than apriori.

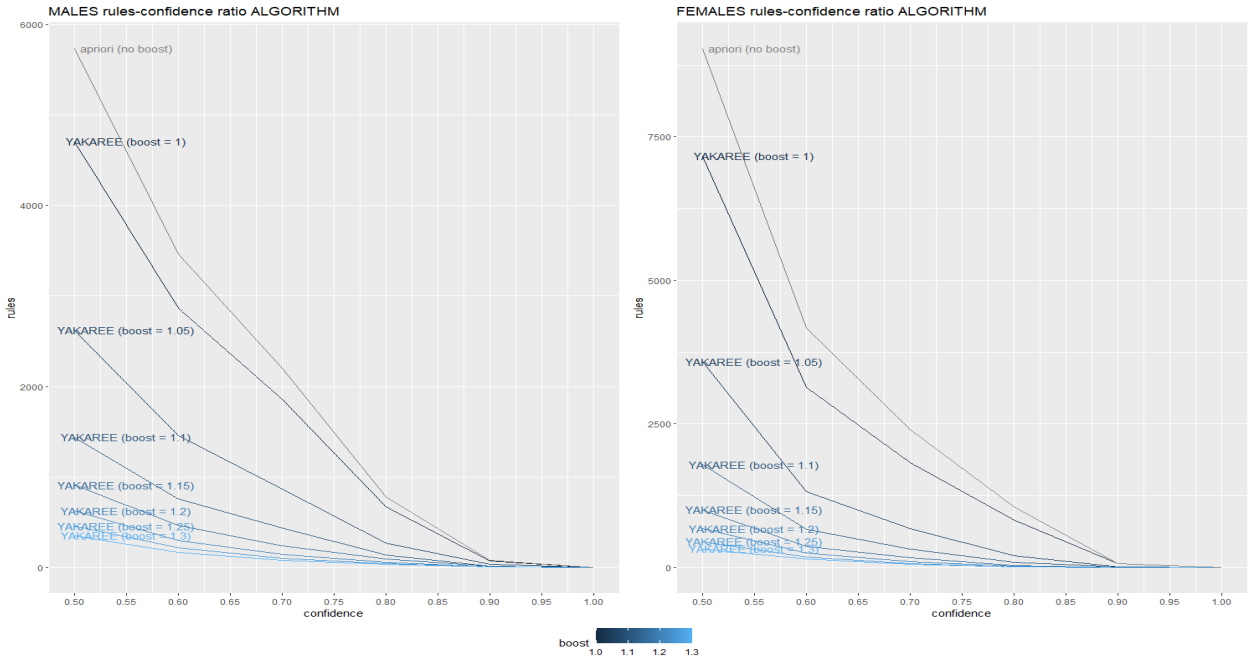

Figure 5.3: **Comparation Yacaree and Apriori rule miners in SEX partitioned datasets.** Comparation of the two algorithms in the overall dataset for different values of confidence and boost thresholds.Outputs the number of rules

.

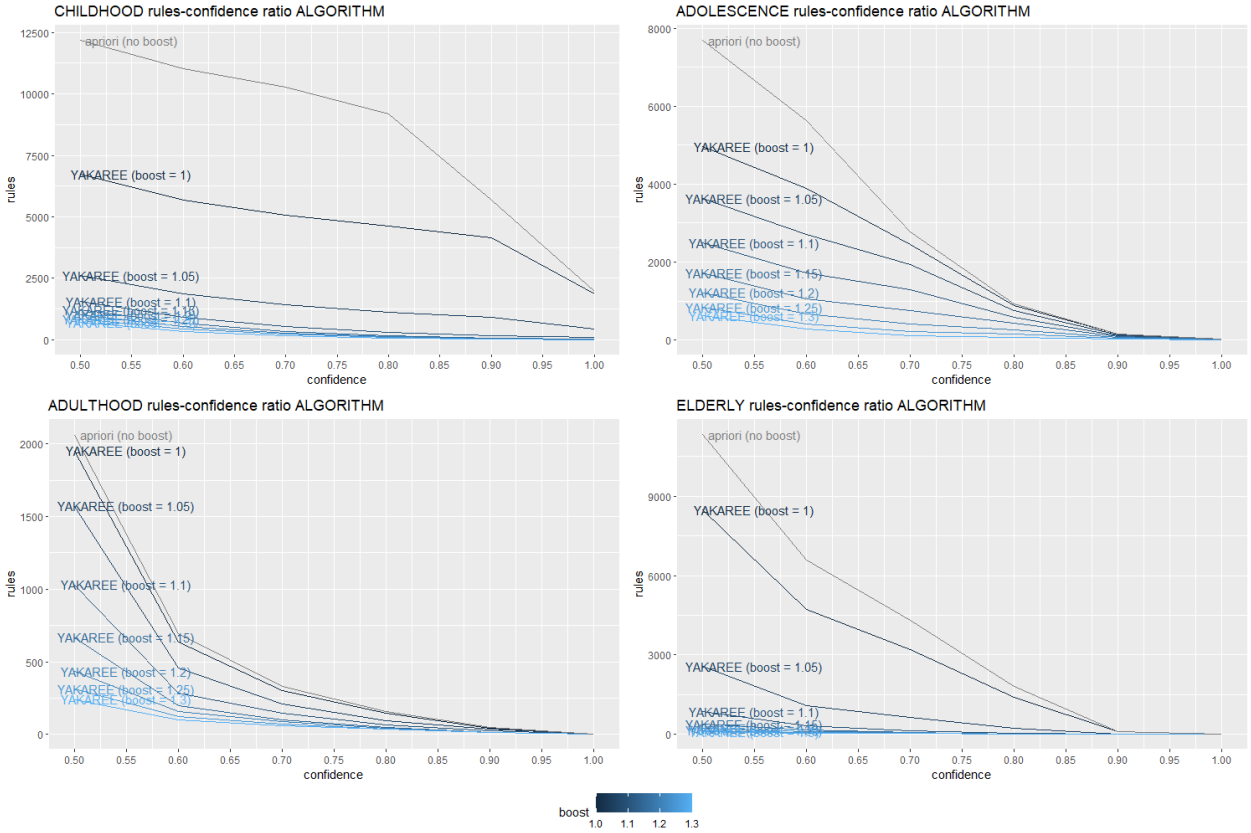

Figure 5.4: **comparison Yacaree and Apriori rule miners in AGES partitioned datasets.** Comparation of the two algorithms in the overall dataset for different values of confidence and boost thresholds.Outputs the number of rules

.

Images 5.2,5.3,5.4 show the results of the different parametrizations in the different datasets used. In all datasets, the yacaree algorithm reduces significantly the number of rules retrieved, avoiding the uninteresting redundant ones. In almost all datasets, regarding the confidence, one can find 'elbows' at confidence thresholds 0.6 and 0.8, which progressively flatten the curves, possible indicatingt good parametrizations to the model. Regarding the confidence boost, at around thresholds 1.15 or 1.20, the number of redundant rules that are being collapsed from the basis drops significantly.

It is interesting the case of the childhood age. Several implications (rules of confidence = 1.0) can be found. Nevertheless, when using the confidence boost, most of them turn to be redundant, reaching 1 single rule for those more demanding boost thresholds.

From the datasets, one could infer that the correct parametrizations are those using a confidence boost between 1.15 or 1.20 and confidence thresholds 0.6 or 0.8, depending on the confidence of the rule wanted. Results of plots comparisons can be found at: yacaree-apriori comparison plots.

5.1.3 Centers of gravity

A center of gravity is the term used to describe those diagnoses or conditions that due to their higher support in a dataset, are overrepresented in the rule output, possibly concealing other important associations in our dataset. The term is defined after the closeness to a center of gravity these conditions have on the basis of a network plotting the rules.

Table 5.7 shows the top-10 rules retrieved by yacaree at confidence boost 1.3 and confidence 90%. As stated in section 5.1 in the case of the raw dataset, the number of trajectories of adults accounts for 52.97% of all the trajectories in the dataset. Moreover, in the case of sex, the trajectories of females account for 54,71% of all trajectories. Thus, this is translated in the rules output. The adult females are over-represented, since the support of these conditions is higher than others, easily passing the support thresholds. Without entering the validity of the rules (since they have passed confidences and confidence support threshold, they are probably correct associations), one must take into account that these over-represented conditions might conceal other conditions such as sex = MALE or the other ages groups. This is the main reason to partition the datasets based on the confounding factors.

Raw dataset	**Partitioned dataset**
{Z30} → {**age_adulthood, sex=F**}	{Z87} → {Z72}
{N92} → {**age_adulthood, sex=F**}	{J21} → {**Z00**}
{Z34} → {**age_adulthood, sex=F**}	{T46} → {I10}
{N91} → {**sex=F**}	{L22} → {**Z00**}
{N95} → {**sex=F**}	{L20,R50} → {**Z00**}
{N60} → {age_elderly}	{J04,R50} → {**Z00**}
{N40, sexe=M} → {age_adulthood, sex=F}	{Z03} → {Z04}
{C61} → {age_elderly, sex=M}	{N18,R80} → {I10}
{L57} → {age_elderly}	{I11,N18} → {I10}
{N50} → {sex=M}	{B09} → {**Z00**}

Table 5.7: Effect of the different *centers of gravity* found on different datasets. *(Left: RAW min-granularity anonimized dataset ; Right: Males dataset min-granularity)*

When focusing on a partitioned dataset, for example, the case of the MALES, which have shown presence in only 3/10 of the rules in the raw dataset, and using the same thresholds (confidence: 90% / confidence-boost: 1.3) it is clearly demonstrated that there exist rules matching this conditions. The functionality of yacaree which allows automatizing the parametrization of the support is the one killing the possibility to present associations for groups that do not have a high presence in the raw dataset.

Although solving the issue of the centers of gravity for the confounding factors by partitioning the dataset, another issue regarding the centers of gravity arises. It can be readily seen that code Z00 is present in most of the rules. As previously stated in section 5, Z00 acts as a hotchpotch for practitioner. Therefore, we have to get rid of these codes having this behavior. Moreover, there are other codes that are not related to a medical condition, yet defining the nature of the visit where they are coded. Table 5.8 shows the codes that have been avoided from now on.

Code	Description (ICD-10)
Z00	Encounter for general examination without complaint, suspected or reported diagnoses
Z01	Encounter for other special examination without complaint, suspected or reported diagnoses
Z02	Encounter for administrative examination
Z03	Encounter for medical observation for suspected diseases and conditions ruled out
Z04	Encounter for examination and observation for other reasons
Z05	Encounter for observation and evaluation of newborn for suspected diseases and conditions ruled out
Z08	Encounter for follow-up examination after completed treatment for malignant neoplasm
Z09	Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm
Z11	Encounter for screening for infectious and parasitic diseases
Z12	Encounter for screening for malignant neoplasms
Z13	Encounter for screening for other diseases and disorders
Z23	Encounter for immunization
Z29	Encounter for other prophylactic measures
Z30	Encounter for contraceptive management
Z31	Encounter for procreative management
Z32	Encounter for pregnancy test and childbirth and childcare instruction
Z34	Encounter for supervision of normal pregnancy
Z36	Encounter for antenatal screening of mother
Z39	Encounter for maternal postpartum care and examination
Z43	Encounter for attention to artificial openings
Z44	Encounter for fitting and adjustment of external prosthetic device
Z45	Encounter for adjustment and management of implanted device
Z46	Encounter for fitting and adjustment of other devices

Table 5.8: **Codes filtrated to improve the results.** The codes relate to hotpoch usually used by the practitioners when misscodify the visits.

Although these codes are uninformative based on our research question of morbidity detection, just by changing this question the method could still be interesting to mine the associations. These codes are informative if we want to know how medical practitioner code or even the resources used by the medical healthcare system, by knowing which codes leads to each visit or medical procedure and the cost it has from an economic point of view.

After filtering the males' dataset, table 5.9 shows the rules found for same parametrization (confidence: 90% and confidence boost: 1.30). Still, hypertension acts as a center of gravity, but since it is a diagnoses coding for a disease, we will keep it. This behavoiour is interesting since may be pointing that if we are interested in a certain disease we can avaluate the associations related to it by segmenting the portion of population undergoing it, treating it as a confounding factor.

Rule	Parameters	Traduction
{Z78} → {Z72}	[conf: 0.911; supp: 0.018; lift: 37.135; boost: 1.683]	{Personal history of other diseases and conditions} → {Problems related to lifestyle}
{T46} → {I10}	[conf: 0.923; supp: 0.004; lift: 3.804; boost: 1.449]	{Poisoning by, adverse effect of and underdosing of agents primarily affecting the cardiovascular system} → {Essential (primary) hypertension}
{N18, R80} → {I10}	[conf: 0.909; supp: 0.005; lift: 3.747; boost: 1.135]	{Chronic kidney disease (CKD), Proteinuria} → {Essential (primary) hypertension}
{I11, N18} → {I10}	[conf: 0.902; supp: 0.004; lift: 3.716; boost: 1.128]	{Hypertensive heart disease, Chronic kidney disease (CKD)} → {Essential (primary) hypertension}
{E79, I11} → {I10}	[conf: 0.930; supp: 0.003; lift: 3.833; boost: 1.120]	{Disorders of purine and pyrimidine metabolism, Hypertensive heart disease} → {Essential (primary) hypertension}
{I11, M10} → {I10}	[conf: 0.900; supp: 0.003; lift: 3.708; boost: 1.084]	{Hypertensive heart disease, Gout} → {Essential (primary) hypertension}
{J00, R19} → {K52}	[conf: 0.903; supp: 0.006; lift: 17.602; boost: 1.081]	{Acute nasopharyngitis [common cold], Other symptoms and signs involving the digestive system and abdomen} → {Other and unspecified noninfective gastroenteritis and colitis}
{J44, Z87} → {Z72}	[conf: 0.985; supp: 0.003; lift: 40.137; boost: 1.081]	{Other chronic obstructive pulmonary disease, Personal history of other diseases and conditions} → {Problems related to lifestyle}

Table 5.9: **Rules MALES Dataset.** Top-10 rules found in males dataset after cleaning the centers of gravities related to uninformative codes in 5.8.

It is also interesting to note the case of the childhood age. From section 5.1 we know that a lot of implications are present in the dataset. When deepening in the output:

RAW (5690 rules)	**CLEANED (1779 rules, -31.26%)**
{} → {**Z00**}	{K04} → {K02}
{J21} → {J00, **Z00**}	{R63} → {**E66**}
{Z91} → {T78, **Z00**}	{K08} → {K02}
{R19} → {K52, **Z00**}	{J21} → {**J00**}
{**Z03**} → {**Z00**, **Z04**}	{B37} → {**J00**}

Table 5.10: **Rules CHILDHOOD Dataset.** Top-5 rules found in males dataset after cleaning the centers of gravity related to uninformative codes in 5.8.

It is noted that all the rules are related to Z codes. Since Z00 has high confidence and support on its own (88,1% of the trajectories present it). Acts as a Gravity center forcing all resulting rules to be created towards them. When correcting the dataset, the number of rules decreases (-31,26%) and the subjective evaluation has higher reliability, giving information about interesting codes.

From the "centers of gravity" intuition, it is easy to check that when a certain dataset is not balanced by the support of all the codes, the code that has higher support tends to be present in a higher number of rules. Nevertheless, having a balanced medical dataset turns to be erroneous most of the time, since there are many variables and we will be missleading important information.

Based on this definition, if a certain condition is wanted to be present in the associations, although not implicitly, one must convert it to a confounding factor, in order to examine all the rules present in this group of individuals. As an example, if we are interested in a certain diagnoses such as 'diabetes mellitus type II', one should create a separated dataset with all the trajectories presenting the condition, erase it from the trajectories (or not, based on if we want the condition to be an implication), and compute the trajectories.

It is important to recall that this procedure will have different usefulness based on the research question tackled:

- If we keep the condition present in the trajectories from the confounding factor dataset (I.e. the dataset of adults having diabetes). We will be finding the associations between diagnoses and the target condition, therefore conditions co-occurring with the illness. Morbidities with the target disease. Since diabetes will be an implication, almost all rules will have it.
- If we don't keep the condition present in the trajectories from the confounding factor dataset (I.e. the dataset of adults having diabetes, avoiding the coding of it). We will be finding the conditions co-occurring in the group of persons that present the disease. Which conditions and morbidities define the group of patients that can undergo diabetes.

5.1.4 Discussion

From this chapters, it is evident that a large number of rules given by classic association miners such as a priori give large fractions of representative rules that are uninteresting, lacking novelty.

By applying a little impact on the confidence boost, the amount of rules passing the thresholds drops rapidly, getting rid of those uninteresting rules. Therefore, from an empirical point of view, the confidence boost clearly improves these results from classical miners such as a priori, by getting rid of the uninformative rules. Nevertheless, it is important to note that at a certain point, even if increasing the confidence boost threshold, the number of uninformative rules decreases slowly, eventually reaching a point where interesting rules are being avoided. It is important to note that although lesser, yacaree also suffers from the expertise of the end-user using the technique

Yacaree could be a good approach to tackle the dataset under study. While in other fields confidence of 60% or 70% could be not of interest, in health terms confidence of 60% for support high enough could affect a large portion of the population, possibly being a good target to study or to attach importance. Sometimes, confidence of 60%, although not being very high, could unearth possible associations between conditions, treatment, sociodemographic groups...

5.2 Visualization

In order to visualize the rules, from a subjective point of view, we need to find a method that can be useful for the practitioner. Traditionality, the inspection of rules is based on 2D plots of confidence-support or confidence-lift, or visualization based on the exploration of the rules variables, not entering on the meaning of the items in rules and its interactions.

By using a network visualization tool, one is able to differentiate between diagnoses and rules and understand the associations from a medical perspective, not only trusting the variables results (lift, support...) that turn to be less intuitive. For the subjective evaluation procedure, we will use the network plotting frameworks in R VisNetwork and iGraph .

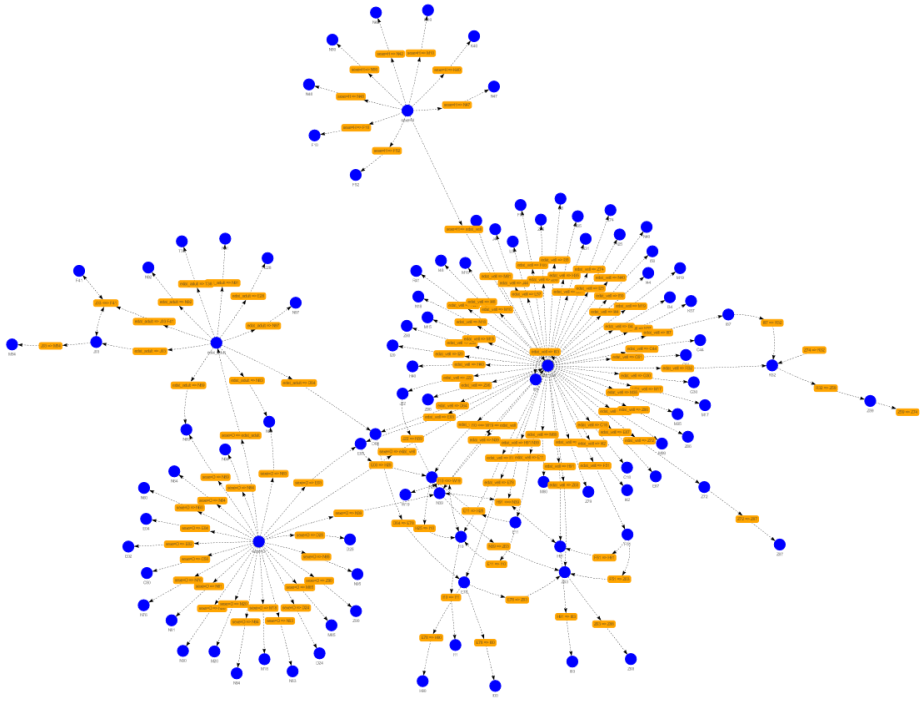

Figure 5.5: **Final network created.**

.

The visualization can be achieved by traducing the output from Yacaree framework to the R class *arules* from arules package, a class devoted to analyze association rules. Next, arulesViz provides several visualization techniques to plot these rules. We will be using the network visualization types. These visualizations will be called *rules networks* in the following sections of the thesis.

Association rules must be interpreted in order to gain knowledge for a certain dataset. These rules can be easily interpreted if their numbers relate to self-explanatory items in a predefined context. In the case of medical datasets, rules relate codes that define medical conditions. These codes may be counter-intuitive to the average people, but the conditions descriptions give a little light for interpretability. Nevertheless, association rule visualization techniques always involve the usage of support-confidence plots or matrix-based plotting procedures that although being informative from an empirical point of view, does not provide information from a subjective point of view.

Networks are a good option to relate diagnoses and be able to provide a good interpretation since the relatedness of terms can be easily outlined and one could still give other information such as the lift, support, or results of techniques such as line width, items dimensions, color...

In order to construct the network, we use igraph R library to extract an initial network from the association rules dataset (we have to adapt the yacaree output to the class arules in R to be able to use all the information). Second, using visNetwork (a more specific library to construct networks) we can improve the visualization. Image 5.6 shows the output of the rules present in the male dataset.

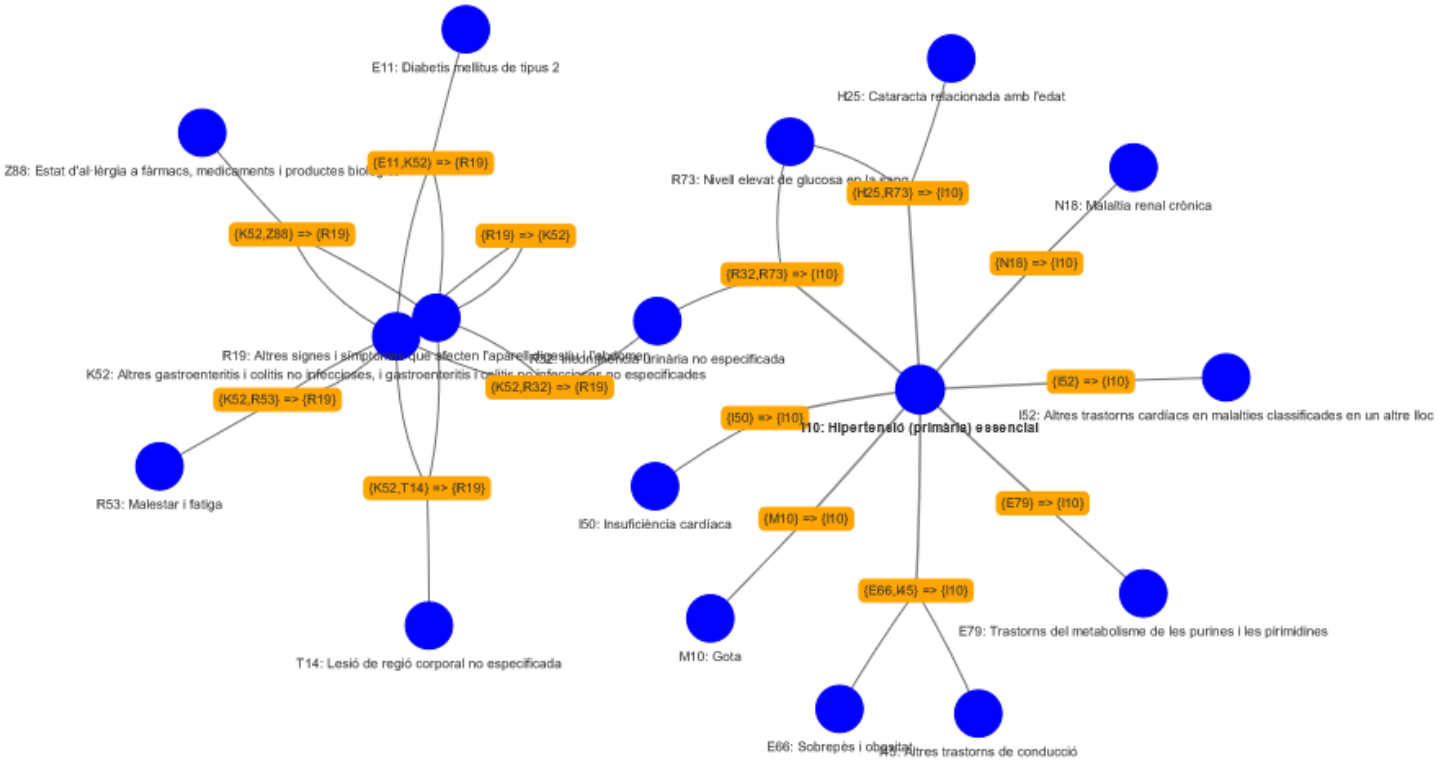

Figure 5.6: **Rules network plotting.** Rules output from FEMALE dataset at confidence-boost 1.20 and confidence 80%.

Definition 12 *The rule network of dataset D defines the associations of diagnoses of patients in a given dataset for confidence γ, support τ, and confidence-boost β.*

Although interesting and useful, this approach suffers for these datasets that are not so clear. In the case of adults and elderly populations, where the datasets are governed by strong association rules presenting high support and confidence, the rules network is clearer. In the other hand, in the case of adolescence and childhood, the amount of codes and rules grows exponentially, with no very clearer results. Images 5.7, 5.8 show the differences in rules networks. Yearly rule networks can be found at: Yearly rule networks.

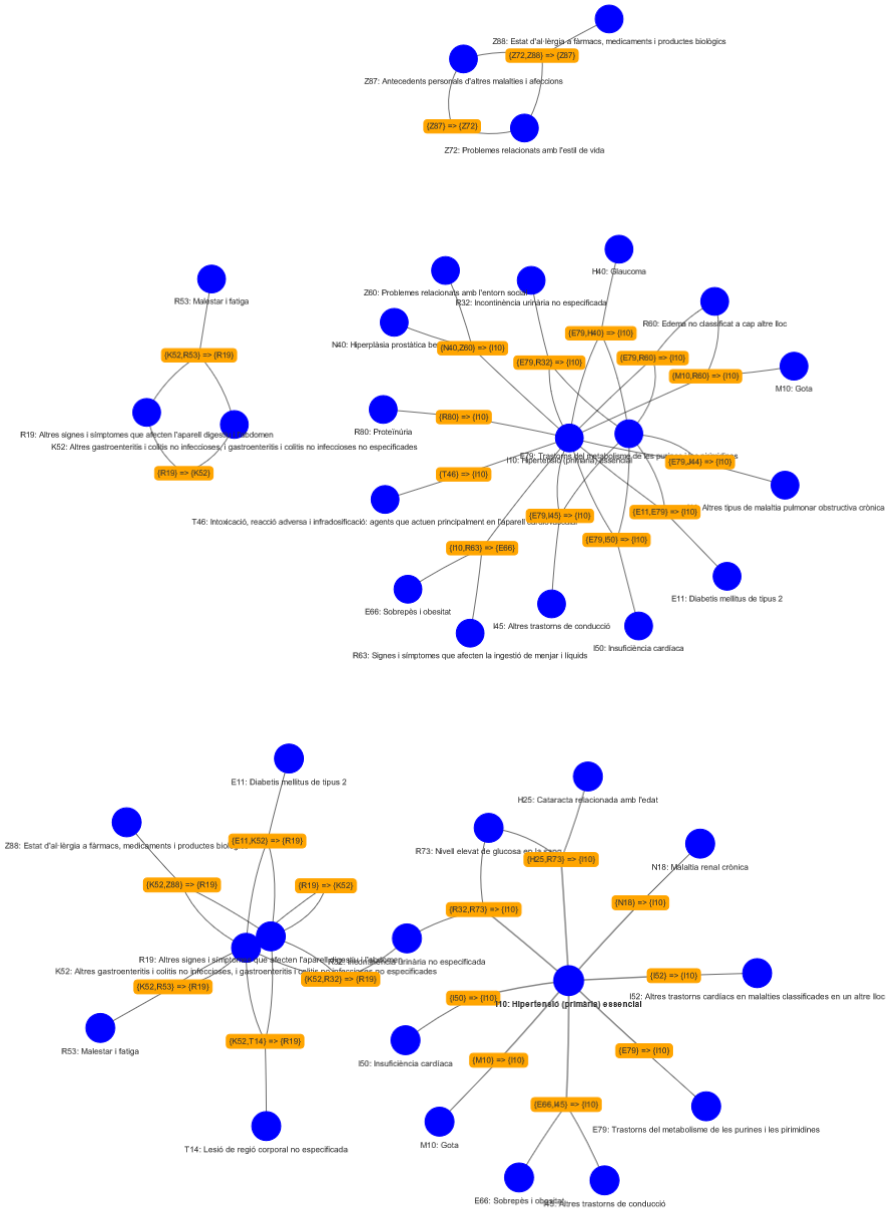

Figure 5.7: **Rules network plotting.** Rules output from SEX partitioned dataset at confidence-boost 1.20 and confidence 80%. *Left: MALES; Right: FEMALES.* Sex rule networks.

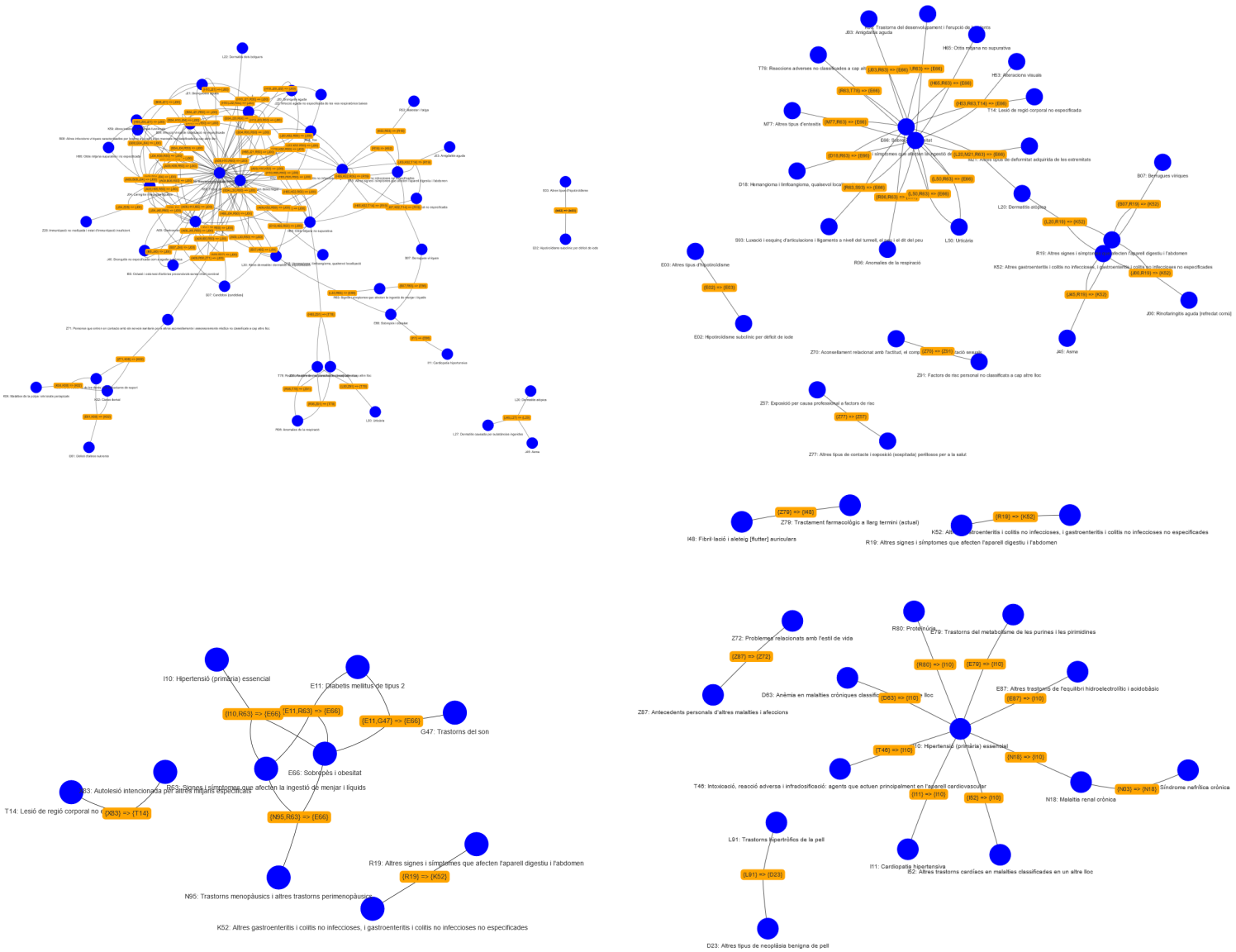

Figure 5.8: **Rules network plotting.** Rules output from AGES partitioned dataset at confidence-boost 1.20 and confidence 80%. *Top-Left: CHILDHOOD; Top-Right: ADOLESCENCE; Bottom-Left: ADULTHOOD; Bottom-Right: ELDERLY.* AGE rule networks.

5.3 Discussion

Confidence boost algorithm has proven to work both empirically and subjectively.

From an empirical point of view, yacaree has succeed at reducing considerably the amount of rules retrieved to a better interpretable subset. Therefore, it is evident from the results that the redundancy notion allows not to loose as much information as when using simply confidence-support framework, since some of this information is being collapsed in a smaller subset of irredundant rules.

Yacaree improves the results when compared to apriori. The algorithm correctly reduces the number of rules to a better irredundant basis of provably minimum size based on the thresholds. Nevertheless, Yacaree cannot compete with the computation efficiency of "apriori" (lasting the last less than ten seconds in all datasets while yacaree lasts nearly 20 minutes in the case of the most costly computing datasets).

Subjective evaluation has provided correct results for the high-demanding thresholds evaluated, supplying correct evident rules that prove the correctness of the model. After cleaning some un-informative codes that act as centers of gravity, the results improve further. The effect of reducing thresholds has to be evaluated in further steps. Some intuitions suggest increasing confidence boost threshold while progressively reducing the confidence, the explore less demanding associations and not suffering from the increasing base size dimension, by controlling the redundancy.

The algorithm suffers from the effect of support. Much of the rules tend to relate a center of gravity. There are codes presenting a higher support, unbalancing the dataset, and pointing the interest of the model towards them, possible avoiding other interesting associations relating codes with smaller support. Results are improved when partitioning the dataset. However, the effect will be always visible since is not feasible to achieve a balanced dataset for all the diagnoses, and will not be a good cleaning procedure. One intuition is to partition the datasets based on a certain condition of interest to achieve a map of associations towards it.

Chapter 6

Morbidity

This chapter of the thesis is devoted to pushing harder the concept of association rule mining in the medical field. Since trajectories are made up from diagnoses one could suppose that rules are the interactions between diagnoses, and therefore illnesses, for a certain group of people. Assuming the algorithm works well, and the dataset is good enough to capture a significant portion of the population under study, one could push the intuition of interaction between illnesses to reach a morbidity suggestion.

This section is intended to suggest groupings between target rules in order to detect different morbidity cases among the population, making a global visualization of the diseases and morbidities that suffers a population segment. We will use the yacaree output to propose an approach to identify morbidity propositions, that can be evaluated by medical experts.

6.1 Context

As stated in previous chapters, in the past some studies have intended to use association rule miners in the 'health data mining context'.

Some articles propose the approach to identify patterns in apnea events [Pombo N. Garcia N. and Bousson K., 2017] or liver disease [Kumar Y. and Sahoo G., 2013], focusing on a specific disease context. In [Lakshmi KS. et al., 2017], association rules are used to mine large volumes of "Electronic Health Records" (EHRs) to find correlations among diseases, symptoms, drugs. Lift is used to discriminate among interesting rules. In [Doddi S. et al., 2001] the goal is to find the relationships between procedures and reported diagnoses in EHR. Association rules are used to extract knowledge in diabetic data repositories in [Stilou S. et al., 2001]. In [Brossette S. et al., 1998] association rules are used to extract interesting patterns in hospital infection control and public health surveillance, focusing on the case of infection of Pseudomonas aeruginosa.

Focusing on the case of comorbidity detection some studies are intended to find comorbidity related to a target disease or group of individuals. In [Fabian P. Held et al., 2016] 17 already known comorbidity were analyzed with several algorithms based on an index disease in elderly men. In [Yueh-Ming et al., 2019], a dataset study of ADHD based on rule mining has been undergone in the National Health Insurance Database of Taiwan. Same in [Wang CH et al., 2019], where the target diagnoses to study the comorbidity are mental disorders. Other studies discover comorbidity and multimorbidity in global EHR databases, not focusing on target diagnoses nor groups. In [Lakshimi KS. and Vadivu G., 2019] a novel approach based on weighted association rule mining is used to discover comorbidity patterns.

6.2 Morbidity

There is no consensus on the meaning of the term "comorbidity" and the related concepts "multimorbidity", "morbidity burden" and "patient complexity" within the medical field. There is an agreement in the fact that it is associated with worst health outcomes, more complex clinical management, and increased health costs. The impact is particularly high in gerontology. When talking about morbidity, the unit of study is the disease, and the term refers to the relation, causality, or directionality between one or more target diseases.

Based on [Valderas JM. et al., 2009] which aims to raise consensus about the meaning of the concepts, morbidity is the term generalizing different concepts that are differentiated based on the constructs and relations under study among the diseases.

- **comorbidity:** "Any distinct additional entity that has existed or may occur during the clinical course of a patient who has the index disease under study". (Needs to be designated an index and a comorbidity condition).

- **Multi-morbidity:** The co-occurrence of multiple chronic or acute diseases and medical conditions within one person. No necessity for a reference index. An example would be dual diagnoses in psychiatry (severe mental illness and substance abuse).

- **Morbidity burden:** Total burden of physiological dysfunction having an impact on an individual's physiological group. It is an index to measure this morbidity in a specific group based on the sex, age or location.

- **Patient complexity:** Extends the morbidity burden not to only the influence of health-related characteristics, but also by socioeconomic, cultural, environmental, and patient behavior.

It is important to note that two comorbid diseases can either be present at the same point of time or occur within a given time period without being simultaneously present. In the later case, is important to note the sequence in which comorbidity appears, which may have important implications for genesis, prognosis and treatment.

For patients presenting depression and diabetes, medically speaking it is not the same to present one or another other of occurrence of the two conditions.

Following sections propose methods to suggest multimorbidity and morbidity burden from the yacaree output procedure.

6.2.1 Patient Complexity

The dataset of trajectories constructed in early section not only contains conditions related to illnesses, but conditions related to factors that can influence the health of a patient. As a consequence, when talking about disease associations, we are also extending patient complexity, since some factors are directly related to the patient's socioeconomic, cultural, environmental, and behavioral conditions. These factors are the codes remaining from ICD-10 chapter: *(Z00-Z99) Factors influencing health status and contact with health services.*

Table 6.1 enhances some of the rules found in subjective evaluation section that might be inferring patient complexity. There are some rules that relate a "personal history of chronic diseases and conditions" with patient complexity:

- From RAW, **rule 1** one could characterize the elderly males with the association between having a *personal history of chronic diseases and conditions* and *problems related to lifestyle.* Although obvious, this proves the correctness of the model.
- From ADULTHOOD, **rule 11** one can see the association between a *long term drug therapy* and *atrial fibrillation and flutter.* Which turns to be also predictable, but demonstrates the correctness.

It must be noted that males present more patient complexity rules than females. The **rule 5** associates:

1. *Problems related to lifestyle.*
2. *Allergy status to drugs, medicament, and biological substances.*
3. *Personal history of other disease sand condition.*

   Could hypothesize the presence of allergies to drugs, medicament, and biological substances in those persons that have had a personal history of diseases, and have taken drugs more often.

The **rule 2** in the RAW dataset relates situations of abuse:

1. *Encounter for mental health services for victim and perpetrator of abuse.*
2. *Problems related to care provider dependency.*
3. *Unspecified urinary incontinence.*

   Possibly pointing the evidence i of "urinary incontinence" in persons that have been abused and present "care dependencies).

There is also a rule that is related to psychological factors, in the ADOLESCENCE group, more prone to suffer them.

- The **rule 10** relates the consultations between *sexual attitude, behavior and orientation* in the healthcare services with *personal risk factors.*

Datasets	**Rule**	**Traduction**
RAW dataset; MALES dataset, ELDERLY dataset	Z87 → Z72	Personal history of other diseases and conditions → Problems related to lifestyle
RAW dataset	{Z69, Z74} → R32	{Encounter for mental health services for victim and perpetrator of abuse, Problems related to care provider dependency} → Unspecified urinary incontinence
	{E11, Z60} → I10	{Type 2 diabetes mellitus, Problems related to social environment} → Essential (primary) hypertension
MALES dataset	{N40, Z60} → I10	{Benign prostatic hyperplasia, Problems related to social environment} → Essential (primary) hypertension
	{Z72, Z88} → Z87	{Problems related to lifestyle, Allergy status to drugs, medicaments and biological substances} → Personal history of other diseases and conditions
CHILDHOOD dataset; ADOLESCENCE dataset	Z77 → Z57	Other contact with and (suspected) exposures hazardous to health → Occupational exposure to risk factors
CHILDHOOD dataset	{Z71, K08} → K02	{Persons encountering health services for other counseling and medical advice, not elsewhere classified, Other disorders of teeth and supporting structures} → Dental caries
	{A09, R50, Z71} → J00	{Infectious gastroenteritis and colitis, unspecified, Fever of other and unknown origin, Persons encountering health services for other counseling and medical advice, not elsewhere classified } → Acute nasopharyngitis [common cold]
CHILDHOOD dataset	{R06, Z91} → T78	{Abnormalities of breathing, Personal risk factors, not elsewhere classified} → Adverse effects, not elsewhere classified
ADOLESCENCE dataset	Z70 → Z91	Counseling related to sexual attitude, behavior and orientation → Personal risk factors, not elsewhere classified
ADULT dataset	Z79 → I48	Long term (current) drug therapy → Atrial fibrillation and flutter

Table 6.1: Rules extending patient complexity in the different datasets mined. *Using confidence-boost 1.20 and confidence 80%.*

From the results, it is evident that the algorithm is able to catch not only interaction between diagnoses and medical conditions, but also interactions between external factors extending patient complexity. Given the definition of the medical datasets coded in *ICD-10/CIM-10*, the resulting outputs will be always suggesting patient complexity.

6.2.2 Multimorbidity

Historically, multimorbidities are a set of coexistence medical conditions. Multimorbidities are associated with poorer medical outcomes and the increased use of health and social care services with the corresponding associated costs. There is an increasing awareness that healthcare services, traditionally single disease-focused, are not correctly designed to meet the challenges of multimorbidity. Practitioners generally face challenges in using medical guidelines designed to treat single conditions or a group of similar conditions. These issues may bring an associated risk such as polypharmacy.

Despite these challenges, there is no international consensus on the best way to define and measure multi-morbidity.

In this section, we aim to propose a technique to detect possible multimorbidity groups. Moreover, we will also use patient complexity data present in codes Z55-Z65 in the study, extending the definition of multimorbidity to also detect multimorbidity groups associated to factors influencing the patients health.

In order to detect those groups, we have performed a hierarchical clustering to the networks retrieved from the yacaree output, to detect communities of related conditions. The method finds modules densely connected themselves but sparsely connected to others in the network and performs a clusterization. Resulting dendograms, morbidity groups and morbidity networks can be found at: morbidity assesment.

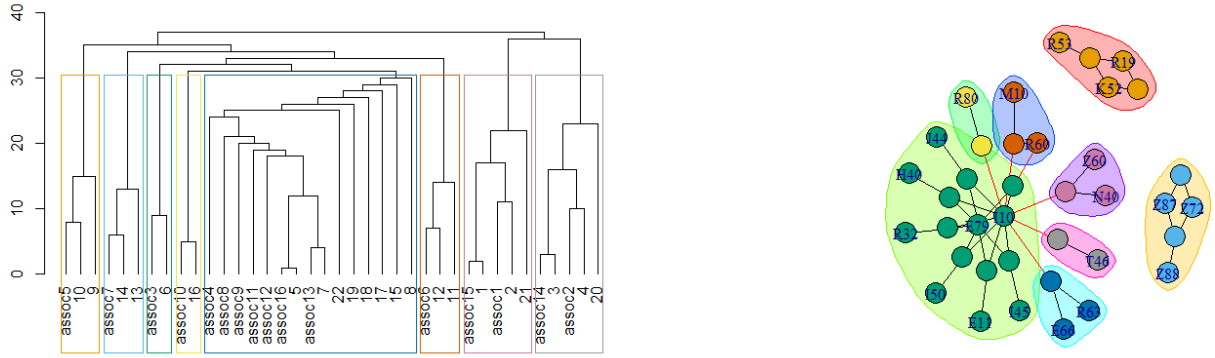

Figure 6.1: **Multimorbidity groups detection.** Procedure to detect the multimorbidity groups in ADULT dataset. Based in Hierarchical clustering.

The method does not use data from rules such as support, confidence or lift. In further steps, this data could be used to leverage the different edges to improve the community detection.

Group	diagnoses
A	I10: Essential (primary) hipertension. I52: Other heart disorders in diseases classified elsewhere. D63: Anemia in chronic diseases classified elsewhere. E79: Disorders of purine and pyrimidime metabolism. E87: Other disorders of fluid, electrolyte and acid-base balance.
B	I10: Essential (primary) hipertension. N03: Chronic nephritic syndrome. N18: Chronic kidney disease.
C	I10: Essential (primary) hipertension. R80: Proteinuria.
D	I10: Essential (primary) hipertension. T46: Poisoning by, adverse effect of underdosing of agents primarly affecting the cardiovascular system.
E	I10: Essential (primary) hipertesnsion. I11: Hypertensive health disease.
F	D23: Other beningn neoplasms of skin. L91: Hyperthrophic disorders of skin.
G	R19: Other symptoms and signs involving digestive system and abdomen. K52: Other and unspecified noninfective gastroenteritis and colitis.
H	I48: Atrial fibrillation and flutter. Z79: Long term (current) long therapy.
I	Z72: Problems related to lifestyle. Z87: Personal history of other diseases and conditions-

Table 6.2: **Morbidity groups discovered.** Morbidity groups inferred from the MALES dataset at confidence boost 1.20 and confidence 80%.

The procedure shows promising results. Since as a validation we are using high demanding confidence thresholds, the morbidities suggested tend to be obvious, with some of them relating very similar conditions. To assess multimorbidity several research studies must be held in parallel. Nevertheless, the algorithm works well at identifying groups of similar conditions and interactions, thus identifying groups of related conditions.

From table 6.2, some intuitions regarding the multimorbid groups can be extracted:

- Groups A,B,C,D,E relate several disseases with hipertension. While A and B are multimorbid groups (presenting more than two diseases), groups C,D and E could be redefined as comorbidity conditions to hipertension target condition.
- Group A relates CARDIOVASCULAR conditions.
- Group B relates NEPHRITIC SYSTEM conditions with hypertension.
- Group C defines comorbidity between hipertension and a disease of the urinary system.
- Group D defines comorbidity between hipertension and underdosing of some drugs related to cardiovascular system.
- Group E defines comorbidity between HIPERTENSIVE conditions.
- Group F defines comorbidity between skin CANCER.
- Group G defines comorbidity between DIGESTIVE conditions.
- Group H defines comorbidity between CARDIOVASCULAR diseases.
- Group I defines comorbidity between PATIENT COMPLEXITY factors.

6.3 Discussion

Yacaree has proven to detect correctly the associations between groups of patients and codes relating the patient complexity. The algorithm detects from a macro point of view, which *patient complexity diagnoses* are associated with a certain population segment.

The algorithm does not infer multimorbidity. Nevertheless, by using clusterization algorithms, the procedure is able to suggest groups of similar associations of conditions correctly. By using network visualization techniques, the procedure could take rid of all those methodologies implemented in this field.

From an intuitive point of view, rule miners are morbidity detection algorithms by themselves. When using hierarchical clusterization we are detecting groups of possible related diagnoses. The algorithm correctly detects groups of similar diagnoses. Further research is needed to evaluate the validity and improvements of morbidity group detection.

The procedure could be improved if leveraging the edges with parameters such as the support, confidence or lift prior to its clusterization.

Chapter 7

Directionality detection

The following section is devoted to set a basic simplified approximation to directionality detection in the diagnoses comprising association rules.

It is worth noting to say that the following section implements a simplified approximation to a problem that turns to be highly though mathematically speaking, where come into play highly demanding fields such as *Proccess mining*, *bayes networks* and *Sequence mining*. Therefore, the following section will only propose a simplified, yet basic approximation to directionality detection between diagnoses based on intuitions, since a complete academic development of the problem turns to be out of scope for this TFM thesis, possible filling more than one doctoral thesis due to the demanding fields and problem being tackled.

7.1 Context

The leading scientist Judea Pearl, subsummes in [Pearl J., 2000] its research in causality, turning to be one of the best references for the field.

The book states:
" Causal imagination enables to do many things more efficiently, through a tricky process we call "planning." [...] There are at least three distinct levels of causality: seeing, doing, and imagining.The first cognitive ability, seeing or observation, is the detection of regularities in our environment. The second ability, doing, stands for predicting the effect(s) of deliberate alterations of the environment, and choosing among these alterations to produce a desired outcome. And the third ability, the understanding or imagining. "

Based on these definitions, causality detection helps to subsume information from the world in order to gain knowledge that can be used to improve the efficiency of a process. The author, points that the causality gathers three different distinct levels.

The first level, following our diagnoses case could be identified as **association**, being the same process automatized by the association mining task performed in previous chapters. This section is related to observation of regularities and can be exemplified by the example question: *What does a symptom tell me about a disease?*.

The second causality level is related to **intervention**: *If I take an aspirin, will my headache be cured?*.

The third level is related to **Understanding**: *Was the aspirin what stopped my headache? Why?*.

These causality detection processes are strongly related with *Process mining*. These techniques usually used in Business context aim to analyse operational processes based on event logs to turn data into insights and actions. The objective is to show what people, machines or entities are doing. Moreover, the field is also strongly related with *Secuence mining*, techniques that aim to bring to light frequent subsecuence by using algorithms able to cope with the combinatorial and exponential search space.

All these research lines are closely linked with *bayesian networks*. Probabilistic graphical models using acyclic graphs representing variables and their conditional dependencies.

Some works such as [A Onisko,. et al 1999] aim to create bayesian network models for diagnoses of liver disorders. Other studies aim to create a clinical bayesian network using EHR in [Y Shen et al., 2018; S McLachlan et al., 2020]. Process mining is used to find causality networks of diagnoses in [M Bozkaya et al,. 2009] or sequence mining for diagnosing a certain condition in [Mazarbhuiya and Alzahrani, 2020]

7.2 Proposal

It must be noted that when plotting rules in a network format, sometimes the interpretation is not straightforward, the increasing number of edges found when the number of output rule increases, as well as the format of the rules, makes the network plots missleading in some cases, prone to miss-interpretation and inducing interesting associations. Although relating terms that are associated based on their co-occurrences, the method lacks at detecting the directionality of associations, non-defining the order of appearance of the diagnoses.

Due to the natural form of the rules network, with some diagnoses acting as centers of gravity (described in section 5.1.3), and by means of support thresholds, most of the rules in the lattice share common diagnoses. Therefore, most of the diagnoses present several edges relating to several rules, which makes the interpretation from a medical point of view difficult.

Supplying the order of appearance of diagnoses in each rule unleashes the initial visualization of rules in a network. This unleashing would not only improve the visualization and the consequent interpretation by avoiding the shared edges between rules presenting the same diagnoses, but also by suggesting directionality, inferring *causality or co-occurrence between diagnoses*. The output would evolve from a visualization indicating major associations between diagnoses in a medical group of patients, to a visualization "suggesting" medical causality paths of diagnoses for a group of patients.

From now on, the output networks will be not relating rules, and therefore association between diagnoses. They will be relating diagnoses and their order. Moreover, the graph will be *directed*, suggesting co-occurrence or causality between diagnoses. In contraposition to definition 12 the directionality network relates diagnoses, without taking into account the initial rules from which they are extracted. Therefore, the irredundant rules found for a certain thresholds are merged to identify the different paths in the lattice.

Definition 13 *The directionality network of dataset D suggests the order of occurence of diagnoses of patients in a given dataset that have been extracted from the rules at confidence γ, support τ and confidence-boost β.*

The idea of the creation of the *directionality network* is to create a directed graph that acts as a process mining network comprising those conditions that have proven to be related suggesting an order of appearance throughout the life of patients. As aforementioned, we will collapse the population database in a graph that can be used as a "guide" for individual cases, recalling in provability based in a simpler basic approximation to directionality.

7.3 Algorithm

Based on the initial intuitions found in [Martí and Gavaldà, 2017] presenting a similar approach to "indicate" directionality, the procedure tests for the significance of the order of appearances between pairs of diagnoses.

Having two conditions A and B, the algorithm checks with a binomial test if the null hypothesis can be rejected. The binomial test has formula:

- H^0*: There is no evidence that* A *or* B *comes first to one another, therefore* AB *and* BA *must have a probability 50%.*

$$p = \sum_{i \in I} Pr(X = i) = \sum_{i \in I} \binom{n}{i} p^i (1-p)^{n-1}$$

Being i the number of successes for a certain "event" and n the total number of events.

The algorithm applies a Bonferroni correction if the appearances of both terms fall below a certain support threshold, making the procedure more robust. A p-value of 0.025 is used to make the model more restrictive, only suggesting a direction if there is a strong evidence.

If the algorithm does not find sufficient evidence against the null hypothesis, the algorithm outputs as a result: '$(AB)'$, meaning there seems not to be causality in the relation, therefore the output edge will be bidirectional. If otherwise, the algorithm outputs a significant order, the output edge will be unidirectional in the output network.

Although the procedure is simple, problems arise when the degree of the rules is higher. For a rule of degree three, with conditions A, B, C the algorithm tests for combinations AB, AC and BC. If the output is significant for the three tests, there should be a defined order that is correct in most of the cases. Based on the counting of the trajectories presenting orders: ABC, ACB, BAC, BCA, CAB, CBA. The algorithm uses the one with a higher number of outcomes (therefore the significant one) to output the directed edges.

If otherwise, not all the tests turn to be significant, the algorithm tests for $(AB)C$, $(AC)B$, and $(BC)A$. Meaning part between parenthesis does not need to follow a specific order, but is significant that happens prior or after the part outside the parenthesis. The algorithm repeats the procedure as before and either outputs an order such as $C(AB)$ or goes to the last step, which is always significant since tests for order (ABC), which turns to be the same number as n (number of total trajectories presenting $A/B/C$).

As one can easily see, the complexity grows exponentially since the number of permutations that are tested grow with the increasing degree of the rules.

As suspected from the explanation of the procedure the algorithm uses two data structures at a time. The first data structure stores the support of each permutation possible in the dataset D. This data structure is used to select the order presenting higher support if the tests are signification.

In the other hand, another data structure tests the support of the combinations of items. Is important to differentiate the combinations from the permutations, in the last the order is taken into account. Since binomial test is bidirectional it turns to be the same to test $\{A, (BC)\}$ than $\{(BC), A\}$.

To output correctly the elements in a network, the algorithm defines *intermediate states* between conditions that relate 2 items. For interpretability, when writing associations between diagnoses we will define associations between brackets and diagnoses outside them. Table 7.1 shows the grammar when writing cardinalities in the format of the desired network.

directionality sintaxis	Network sintaxis
[AB]	$A \rightarrow \{A \Rightarrow B\} \rightarrow B$
(AB)	$A \leftrightarrow \{A \Leftrightarrow B\} \leftrightarrow B$
[AB](C)	$(A \rightarrow \{A \Rightarrow B\} \rightarrow B) \leftrightarrow \{A, B \Leftrightarrow C\} \leftrightarrow C$
(ABC)	$A \leftrightarrow /B \leftrightarrow /C \leftrightarrow \{A \Leftrightarrow B \Leftrightarrow C\}$

Table 7.1: **Directionality traduction.** Traduction from the output format to the networking writing format.

This *intermediate states* are no rules, although extracted from them. These states are used to store the variables such as the support, confidence, and lift of an association.

A second algorithm is devoted to making the translation from the *directionality sintaxis* to the *network sintaxis*.

Figure 7.1: **Directionality example.** Example of the directionality computation of the network of MALES. *Rule network.*

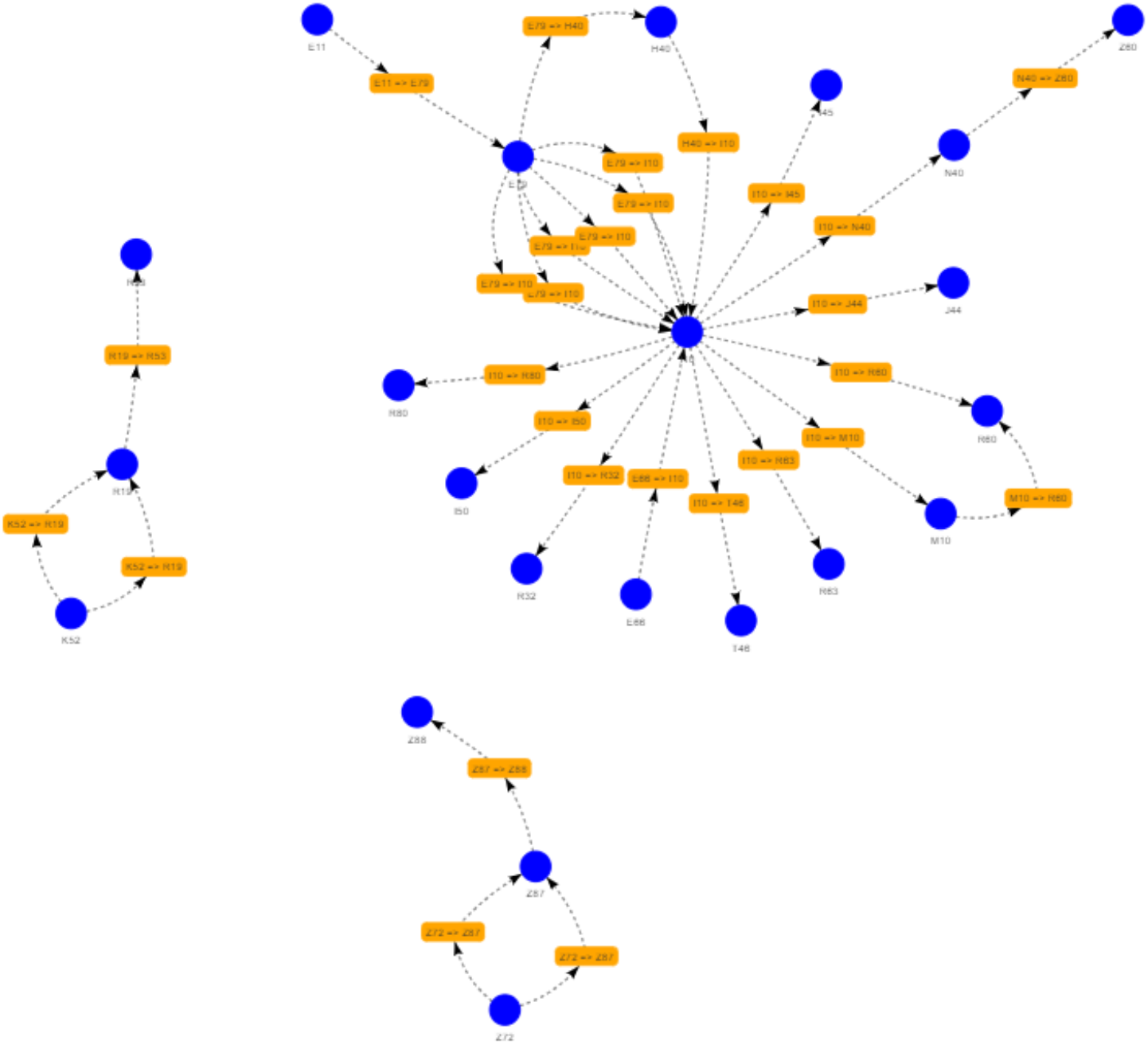

Figure 7.2: **Directionality example.** Example of the directionality computation of the network of MALES. *Directionality network.*

Example 7 *We have performed an association rule mining using Yacaree with parameters: confidence boost: 1.25; confidence: 0.80 and the obtained rules are:*

1. $M10 \rightarrow I10$
2. $\{I16, Z63\} \rightarrow R32$
3. $\{E66, W19\} \rightarrow I10$

Associating the diagnoses:

- *M10: Gout.*
- *I10: Essential (primary) hypertension.*
- *I16: Hypertensive crisis.*
- *Z63: Other problems related to primary support group, including family circumstances.*
- *R32: Unespecified urinary incontinence.*
- *E66: Overweight and obesity. W19: Unespecified fall.*

We want to obtain the graph suggesting directionality from the three rules.

Initially, rules are sorted in increasing order of degree. The simpler rules will be evaluated first. In the first round, only rule 1 will be evaluated. It is important to note that both LHS and RHS are merged in a single subset of diagnoses. From now on, the diagnoses is identified with an id based on their order in the rule, being those ids's the first capital letters of the alphabet. In this case $M10 = A$ and $I10 = B$. The identification is independent of one rule to another. Therefore, in rule 3 I10 will be C, despite being another letter in rule 1.

*The algorithm first constructs the **permutation table**. The possible permutations are AB and BA. The algorithm computes the support of each permutation by searching the number of trajectories presenting both diagnoses in the desired order and dividing it by the total number of trajectories presenting both diagnoses, n.*

*Second, the algorithm computes the **combinations table**. This table is devoted to the testing. Since binomial testing checks for the significance between two options, it will result in the same to test permutation AB or BA, since $AB = p, BA = (n - p)$. With this proposition, the algorithm will test the first alphabetically order permutation, treating it as a combination. It is important to note that (AB) is also a combination to test, is the same as the support.*

After constructing the data structures, the algorithm uses the binomial testing to test the combinations in combinations table. If the number of rules presenting both diagnoses falls below a certain threshold (intuitively set to 500) the algorithm performs a bonferroni correction to make the test more robust.

*Finally, the algorithm checks if the tests are significant ($p - value \leq 0.025$). If the test is significant, there should be an order of appearance clear from the context. The algorithm checks in the **permutation table** the order presenting a higher support and outputs a directionality based on it. If there is not significance, the algorithm checks the second round of tests iteratively until reaching the (AB) test, which always turns to be significant.*

The resulting combinations table:

Actors	**[AB]**	**(AB)**	**[AB]_pval**	**(AB)_pval**
M10:A; I10:B	0.0305	1	1.35e-38	8.55e-50

The resulting permutations table:

Actors	AB	BA
M10:A; I10:B	0.0305	0.969

It is clear from the context ans significant that the good directionality of the association is $[BA]$*, therefore:* $B \rightarrow \{B \Rightarrow A\} \rightarrow A$*. Which is the same to say that I10: Essential (primary) hypertension is present before M10: Gout most of the time, possibly pointing out a causality relation.*

At this point and prior to get to the next degree, the algorithm would store in a separate data structure the results of the actors and the directionality. In the posterior steps, the algorithm would check prior to starting testing if the result of an association is already computed.

In the next iterations, focusing on rules 1 and 2 the results would be:

Finally, the algorithm checks if the tests are significant $(p-value \leq 0.025)$*. If the test is significant, there must be an order of appearance clear from the context. The algorithm checks in the* ***permutation table*** *the order presenting higher support and outputs a directionality based on it. If it is not significant, the algorithm checks the second round of tests iteratively until reaching the* (AB) *test, which always turns to be significant.*

	permutation results						
Actors	[AB]	[AC]	[BC]	[AB](C)	[AC](B)	[BC](A)	(ABC)
I16:A; Z63:B; R32:C	1	0.976	0	0	1	0	1
E66:A; W19:B; I10:C	0.64	1	0.52	0.52	0.84	0.64	1
	p-values						
Actors	[AB]	[AC]	[BC]	[AB](C)	[AC](B)	[BC](A)	(ABC)
I16:A; Z63:B; R32:C	2.3e-84	6.99e-71	2.3e-84	2.3e-84	2.3e-84	2.3e-84	8.0e-87
E66:A; W19:B; I10:C	3.85e-34	0	0.103	0.103	4.39e-34	3.85e-34	0

Table 7.2: **Permutation-test table.** Table of the permutations tested in the example rules.

	Combination results					
Actors	ABC	ACB	BAC	BCA	CAB	CBA
I16:A; Z63:B; R32:C	0	1680	0	0	42	0
E66:A; W19:B; I10:C	300	912	684	0	0	0

Table 7.3: **Combinations table.** Table of the amount of transactions containing the desired combination.

In rule 2 in example 7 the directionality computation is straightforward. Since the three tests are significant in the first testing round, there must exist a defined order with sufficient significant probabilistic robustness. This order is $[ACB]$, therefore $A \rightarrow \{A \Rightarrow C\} \rightarrow C \rightarrow \{C \Rightarrow B\} \rightarrow B$. Which is the same to say that *I16: Hypertensive crisis* is present before *R32: Unspecified urinary incontinence* which turns to be present before *Z63: Other problems related to primary support group, including family circumstances.*

Rule 3 in example 7 is more difficult. In the first testing round two orders have shown evidence to be significant. These are AB and AC. Since not all the codes are significant we enter the second testing round for those subsets that have shown significance. This level tests the amount of times the portion inside brackets [] in the first test is jointly found. Therefore, if wanted to test $[AC](B)$ the algorithm tests $ACB + BAC + BCA + CAB = 912 + 684 = 1596$.

At the second testing round the test $[AB](C)$ is not significant while the other $[AC](B)$ is still significant and thus the correct one. Which is the same to say $(A \rightarrow \{A \Rightarrow C\} \rightarrow C) \leftrightarrow \{A, C \Leftrightarrow B\} \leftrightarrow B$ or that while *E66: Overweight and obesity.* is always found prior to *I10: Essential (primary) hypertension.* the third diagnostic *W19: Unspecified fall.* does not have a clear position in the trajectory regarding the three elements. This turns to be evident, and thus proves the correctness of the model.

7.3.1 Parallelization

The prior computations described in subsection 7.3 are the most costly operations in the overall procedure. Since the algorithm "suggests" the directionality for all rules of a certain degree at a time, the testing can be split in different machines for each degree or storing the event pairs in disk and doing several passes through the sequences databases updating a subset of the event pairs each time.

It is important to note that since most of the rules share common diagnoses and therefore associations, one could store the already computed associations in a different data structure. Before computing these support and probability tests for two items in a rule, the algorithm can "fish" the already computed results, leading in resources saving and a boosting of the procedure.

7.3.2 Prunning

After the preprocessing, since some rules share common diagnoses, several edges turn to be equal. These same edges are pruned in order to only show single relation, since are redundant. Figure 7.3 shows the procedure.

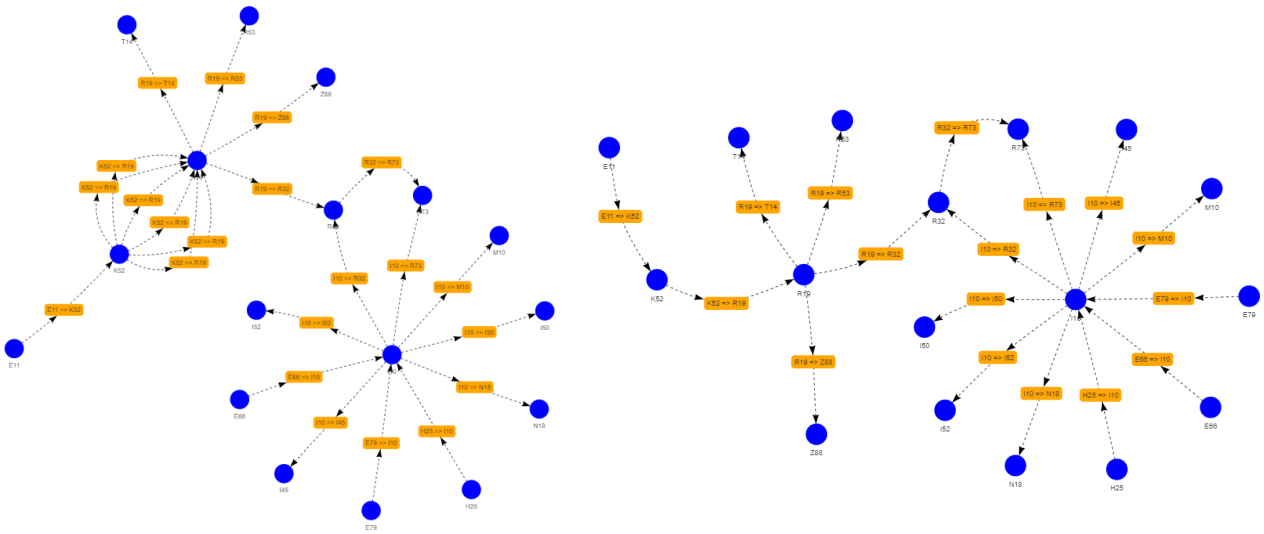

Figure 7.3: **Prunning example.** Example of the prunning of the network of FEMALES. *Left: prior to prunning; Right: After pruning.*

In [Martí and Gavaldà, 2017] a second pruning step can be performed based on the lift of the association. The lift measures the relation of two conditions. A lift greater than 1 indicates a strong relation between a pair of events while a lift nearer to 1 indicates that two elements does not tend to appear together more than what is expected due to random chance. In our study, we will not prune more the tree, since a prior pruning has been performed in the association rule mining task.

It is important to remember that at this point we are following paths of diseases that are associated, not rules. Therefore, we have unleashed the initial rule network to be able to "suggest" paths rather than associations by themselves. It is noticeable, that not all diagnoses are present in these flowcharts, since only those diagnoses proven to be associated are retrieved from yacaree framework. Next, table 7.4 provide some paths present in the datasets that are though to be informative. If more information is needed whole networks can be found at:

It must be highlighted that childhood present a higher number of rules, presenting a high degree and with more undefined indicated directionality. This is due to the fact that children often go to the healthcare services and present some conditions, that although nonhazardous, almost all children have at some point in its life (conditions such as bronchitis, common cold, nasopharyngitis, sprained ankle). The same thing happens in smaller extend to the adolescence community. These codes, although informative, mask other more hazardous conditions. If wanted more specificity in hazardous conditions, we would have to clean from the dataset those codes that are passive (extending lower through live, non-expecting chronicity).

Resulting final yearly directionality networks can e found at: , SEX directionality networks, AGES directionality networks.

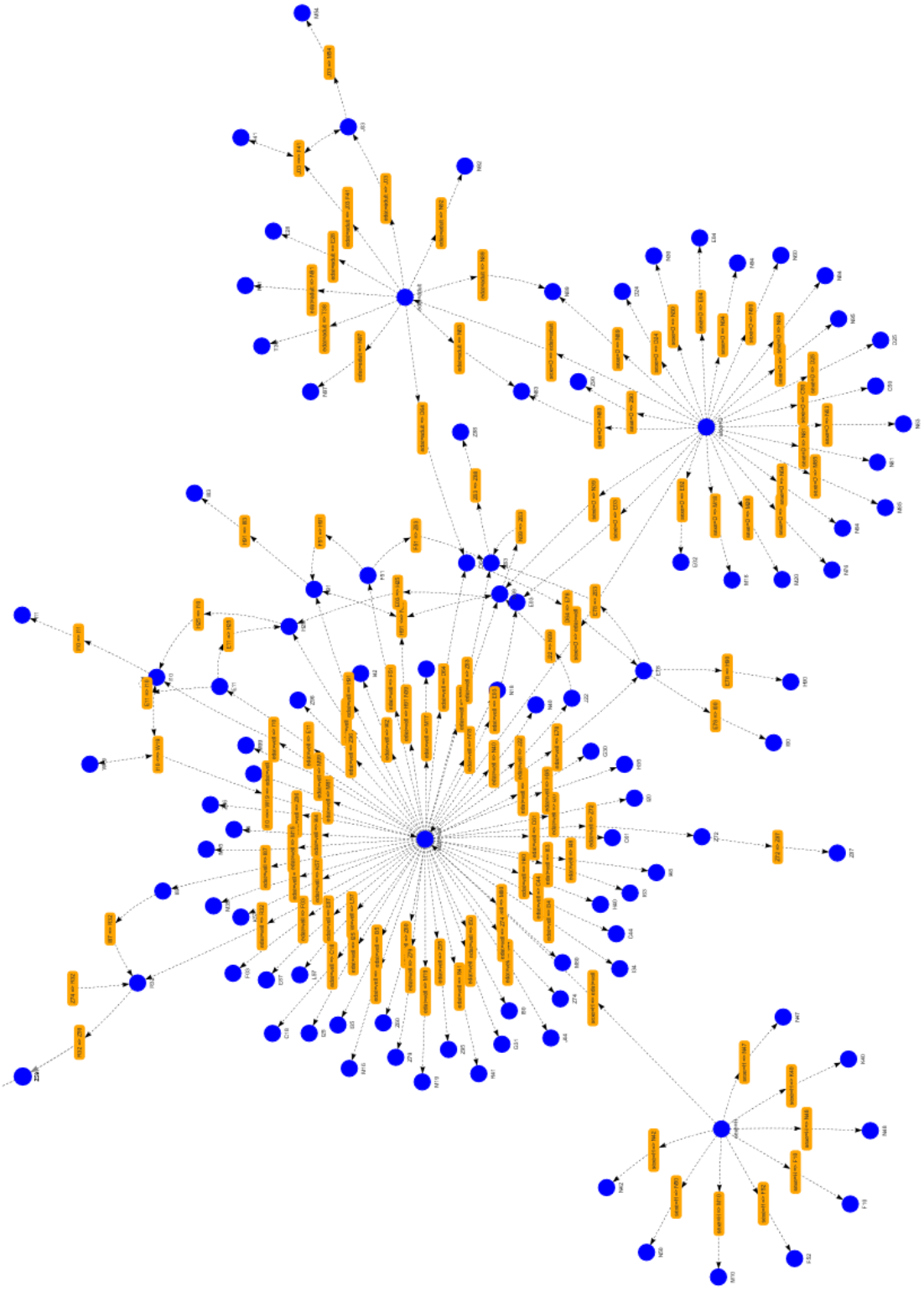

Figure 7.4: **Directionality network RAW dataset 2019.** Directionality network of the dataset from year 2019. Centers of gravity correspond to sexes and ages. RAW 2019 directionality network.

Dataset	Path	Traduction
Males	{K52} ⇒ {R19} ⇒ {R53}	{Other and unspecified noninfective gastroenteritis and colitis} ⇒ {Other symptoms and signs involving the digestive system and abdomen} ⇒ {Malaise and fatigue}
	{E11} ⇒ {E79} ⇒ {H40} ⇒ {H40}	{Type 2 diabetes mellitus} ⇒ {Disorders of purine and pyrimidine metabolism} ⇒ {Glaucoma} ⇒ {Essential (primary) hypertension}
Females	{E66} ⇒ {I10} ⇒ {I50}	{Overweight and obesity} ⇒ {Essential (primary) hypertension} ⇒ {Heart failure}
Elderly	{I48} ⇒ {Z79}	{Atrial fibrillation and flutter} ⇒ {Long term (current) drug therapy}
	{I10} ⇒ {N18} ⇐ {N03}	{Essential (primary) hypertension} ⇒ {Chronic kidney disease (CKD)} {Chronic nephritic syndrome}
Adulthood	{T14} ⇒ {X83}	{Injury of unspecified body region} ⇒ {Intentional self-harm by other specified means}
	{E66} ⇒ {N95} ⇒ {R63}	{Overweight and obesity} ⇒ {Menopausal and other perimenopausal disorders} ⇒ {Symptoms and signs concerning food and fluid intake}
Adolescence	[{R63} ⇔ {E66}] ⇒ {T78}	[{Symptoms and signs concerning food and fluid intake} ⇔ {Overweight and obesity}] {Adverse effects, not elsewhere classified}
	{Z70} ⇒ {Z91}	{Counseling related to sexual attitude, behavior and orientation} ⇒ {Personal risk factors, not elsewhere classified}
Childhood	{B08} ⇒ [{J00} ⇔ {J21}]	{Other viral infections characterized by skin and mucous membrane lesions, not elsewhere classified} [{Acute nasopharyngitis [common cold]} ⇔ {Acute bronchiolitis}]
	[{K04} ⇔ {K08} ⇔ {K02}]	[{Diseases of pulp and periapical tissues} ⇔ {Other disorders of teeth and supporting structures} ⇔ {Dental caries}]

Table 7.4: **Directionality paths.** Some interesting paths found in the different used datasets. Found at networks in : Directionality networks .

7.4 Covid effect

The following subsection is devoted to analyze the procedure in the year of the covid-19 (2020) outbreak.

As indicated in section 4, the pandemic outbreak has seriously affected the medical health services. In the same reference area tackled in the study the overall number of visits have suffered a reduction of -1.36% from between 2019 and 2020. Moreover, not only the number of visits have changed, but also the nature of the visits, the increase in telemedicine has reached a maximum, with an increase of +267%, and a reduction of the face-to-face visits -47%. The diagnoses presenting a higher increase are codes related with COVID-19 (codes *Z20-Z29*, 2.540%) as well as codes related to economic and housing problems (*Z55-Z65*, 44.40%). The principal reductions are related to codes from chronic pathologies such as arterial hypertension (*I10-I16*; -32.73%) or diabetes mellitus (*E08-E13*; -21.13%), but also obesity (*E65-E68*; -48.58%) and bodily injuries (*T14*; -33.70%). Visits with mental health related diagnoses codes have decreased, but less than average. Both for children and adolescents and for adults, there was a decrease in consultations for respiratory infections (*J00-J06*; -40.96%). The results show very significant year-on-year variations (in absolute terms, an average of 12%), a sign of the strong shock to the health system.

The aforementioned year-on-year variations may be responsible of some changes in the associations that can be found using confidence boost yacaree algorithm. The differences between two normal years might be missleading, but the COVID-19 outbreak has changed the way the population relates with the medical services, the diagnosing strategies and even the diseases priority. Some studies indicate that in the year 2021, more neoplasms are being found in later stages of development, occuring the same in the case of memtal illnesses. The following procedure intends to create a network visualization framework to analyse the change of associations between the year pre-covid (reference year, 2019) and the year where the outbreak has had a major impact (2020).

The procedure collapses the network rule graphs obtained using yacaree at *confidence* 80% and *confidence boost 1.20* from both years and plotting the network structure from section 5.1 with the *R* frameworks VisNetwork and iGraph. An algorithm evaluates the rules from both years ans associates a different color to them based on the year of presence: 2019 (yellow), 2020 (red) both (orange). A directionality network is computed using the aforementioned procedure from section 7. Figure 7.6 shows the resulting networks from the RAW dataset.

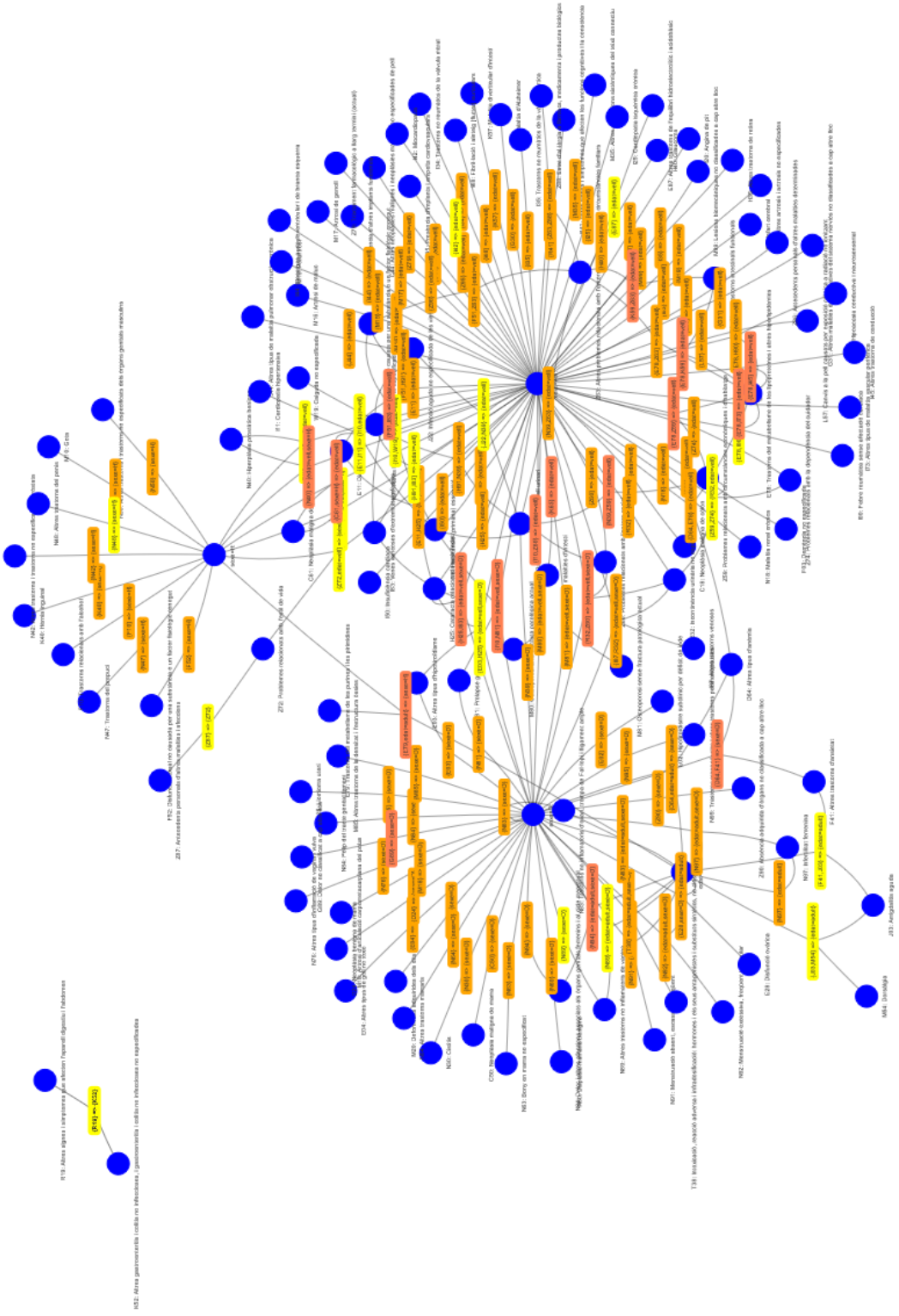

Figure 7.5: **Visualization COV-19 year 2020.** Rules network when adding the year of the COVID-19 outbreak. *Rule network:* *Year comparison Rule networks*

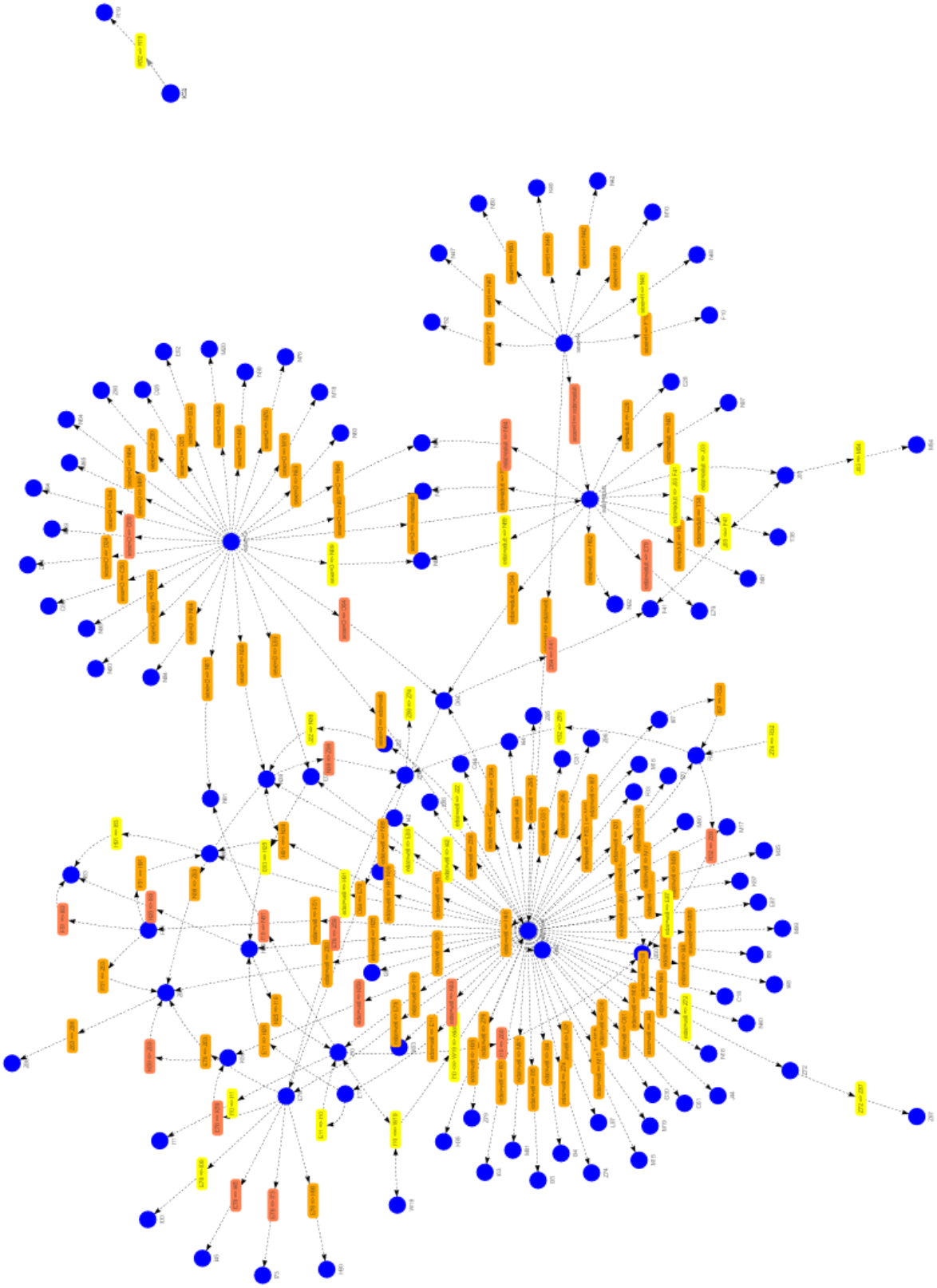

Figure 7.6: **Visualization COV-19 year 2020.** Directionality network when adding the year of the COVID-19 outbreak. *Directionality network: Year comparison directionality networks*

From figure 7.6 is evident that most of the resulting rules and paths are present in both years, 68%. In the other hand the rules only found in 2019 account for a 16% while the ones only present in 2020 account for a 15%. It can be hypothesised that those rules only found in 2019 gather associations that have been infradiagnosed on year 2020, therefore not treated as important. The rules representing only year 2020 are those rules that gather associations diagnoses that have been diagnoses frequently these year, therefore not being infradiagnosed pretty much with respect to others. The rules found in both years are those association that have been string in the two years.

It is important to note that these graphs act over associations, not diagnoses. Therefore we cannot state that a diagnose has been infradiagnosed basing merely on the lack of the association in 2020. All the diagnoses in a rule are treated as a subset, therefore it might be that while diagnoses A, C from rule $A, B \rightarrow C$ have had an equal diagnosing rate in both years, the diagnoses B has decreased below thresholds in these patients undergoing A, C, and the desired rule is not found in year 2020. To identify which diagnose is reponsible from the rule not surpassing the thresholds, database should be inspected for all diagnoses in each rule.

The networks from figure 7.6 are from the RAW dataset, thus directly associating most of the diagnoses with a group of patients. From these directionality networks one is able to see which diagnoses are associated to each group, and how the diagnosing of each age and sex has changed from 2019 to 2020. Table 7.5 shows some of the interesting rules found in the different years associated to each group. Most of the rules are associated to ages *elderly and adulthood* and the both sexes. As mentioned in previous sections of the thesis, the *childhood and adolescence groups*, present more rules with less support, confidence and confidence boost, mainly because don't present strong associations between diagnoses with respect to other ages and account for a smaller number of visits, shifting the association to the *gravity centers* of other ages. The diagnoses found in the different years seem to follow the aforementioned hypothesis of infradiagnosing and importance of the diagnostic.

Table 7.5 turns to match some of the hypothesis stated beforehand. In year 2019 some of the associations between groups and associations are, although evident, from diagnoses that could be defined as non health hazardous. In the other hand, when focusing on those diagnoses present in 2019 and 2020, we found a large number of diagnoses that suppose a health hazard. Examples of these statements are:

- Adult population in 2019 is associated with *tensiolitis* or *anxiety*.
- Elderly population in 2019 is associated with *Varicose veins of lower extremities*, , *Other and unspecified hearing loss*, *Disorders of lipoprotein metabolism and other lipidemias* or *Unspecified urinary incontinence*.
- In 2019 and 2020 elderly and adulthood groups are related with some health hazardous conditions such as *benign mammary dysplasia*, *Malignant neoplasm of colon*, *Other and unspecified malignant neoplasm of skin* or *Other degenerative diseases of nervous system, not elsewhere classified*

Is important to note that although the diagnoses related with neoplasms *Neoplasms (C00-D49* have suffered a drop of -13,57%, there are still a lot of associations related to similar groups between years 2019 and 2020. It is important to note that we are setting a support threshold, therefore, association is not directly detecting the decrease in diagnosing, but detecting that the associations between a condition and a group. Therefore, an association appearing in 2019 and 2020 is telling us that under certain thresholds, the neoplasms have been diagnosed for adult and elderly population in both years, possibly with lower thresholds in 2020 with respect to reference year 2019.

Group	year	Path	Traduction
Raw	2019	{age=Elderly} (⇒ {H91} / ⇒ {E78} / ⇒ {I42} / ⇒ {E03} / ⇒ {W19} / ⇒ {J22})	{ELDERLY} (⇒ {Other and unspecified hearing loss} / ⇒ {Disorders of lipoprotein metabolism and other lipidemias} / ⇒ {Cardiomyopathy} / ⇒ {Other hypothyroidism} / ⇒ {Unspecified fall} / ⇒ {Unspecified acute lower respiratory infection})
		{age=Adulthood} (⇒ {J03} / ⇒ {F41} / ⇒ {N89})	{ADULT} (⇒ {Acute tonsillitis} / ⇒ {Other anxiety disorders} / ⇒ {Other noninflammatory disorders of vagina})
	2019/20	{sex=M} (⇒ {F52} / ⇒ {K40} / ⇒ {N42} / ⇒ {M10} / ⇒ {F10})	{MALE} (⇒ {Sexual dysfunction not due to a substance or known physiological condition} / ⇒ {Inguinal hernia} / ⇒ {Other and unspecified disorders of prostate} / ⇒ {Gout} / ⇒ {Alcohol related disorders})
		{sex=F} (⇒ {N60} / ⇒ {N95} / ⇒ {C60} / ⇒ {N63})	{FEMALE} (⇒ {Benign mammary dysplasia} / ⇒ {Menopausal and other perimenopausal disorders} / ⇒ {Unspecified lump in breast})
		{age=Elderly} (⇒ {M15} / ⇒ {G30} / ⇒ {G31} / ⇒ {N40} / ⇒ {C18} / ⇒ {C44})	{ELDERLY} (⇒ {Polyosteoarthritis} / ⇒ {Alzheimer's disease} / ⇒ {Other degenerative diseases of nervous system, not elsewhere classified} / ⇒ {Benign prostatic hyperplasia} / ⇒ {Malignant neoplasm of colon} / ⇒ {Other and unspecified malignant neoplasm of skin})
		{age=Adulthood} (⇒ {N97} / ⇒ {T38} / ⇒ {N92})	{ADULT} (⇒ {Female infertility} / ⇒ {Poisoning by, adverse effect of and underdosing of hormones and their synthetic substitutes and antagonists, not elsewhere classified} / ⇒ {Excessive, frequent and irregular menstruation})
	2020	{sex=F} (⇒ {G89} / ⇒ {D64})	{FEMALE} (⇒ {Pain, not elsewhere classified} / ⇒ {Other anemias})
		{age=Elderly} (⇒ {K59} / ⇒ {K63})	{ELDERLY} (⇒ {Other functional intestinal disorders} / ⇒ {Other diseases of intestine})
		{age=Adulthood} (⇒ {N94} / ⇒ {E79})	{ADULT} (⇒ {Pain and other conditions associated with female genital organs and menstrual cycle} / ⇒ {Disorders of purine and pyrimidine metabolism})

Table 7.5: **Directionality paths pre/post COVID-19 outbreak.** Some interesting paths found in RAW dataset. RAW yearly comparison directionality networks

Tables 7.6 and 7.7 show interesting association paths found in the newtworks of ages and sex. In the ages of adolescence and adulthood there are several paths related to symptoms associated to covid. These groups of populations are the ones that have been in contact more with the health services, mainly using telemedicine consultation. These results could suggest that although the elderly population have suffered more the effects of covid in the worst forms of the illness, they have not encountered the primary health services as much as others age groups, because the severity of the conditions has shifted the treatment of the elderly population to the leading hospitals, more used to critical clinic scheme. Nevertheless, the adulthood and adolescence groups, could have enccountered more the primary healthcare services, with less critical clinical schemes, leading to more diagnosing of rcovid related diagnoses by the medical professionals. One must remember those groups are more used to technology usage, possible being the major responsibles of the increment in telemedicine.

Group	**year**	**Path**	**Traduction**
Males	2020	{K52} ⇒ {H40} ⇒ {R19}	{Other and unspecified noninfective gastroenteritis and colitis} ⇒ {Glaucoma} ⇒ {Other symptoms and signs involving the digestive system and abdomen}
		{I10} ⇒ {Z79}	{Essential (primary) hypertension} ⇒ {Long term (current) drug therapy}
	2019/20	{I10} ⇒ {T46}	{Essential (primary) hypertension} ⇒ {Poisoning by, adverse effect of and underdosing of agents primarily affecting the cardiovascular system}
		{E79} ⇒ {I10}	{Disorders of purine and pyrimidine metabolism} ⇒ {Essential (primary) hypertension}
Females	2020	{E11} ⇒ {E66} ⇒ {R63}	{Type 2 diabetes mellitus} ⇒ {Overweight and obesity} ⇒ {Symptoms and signs concerning food and fluid intake}
		{I10} ⇒ {I11} ⇒ {N18}	{Essential (primary) hypertension} ⇒ {Hypertensive heart disease} ⇒ {Chronic kidney disease (CKD)}

Table 7.6: **Directionality paths pre/post COVID-19 outbreak.** Some interesting paths found in MALE and FEMALE dataset. SEX yearly comparison directionality networks.

Group	year	Path	Traduction
Childhood	2020	{E66} ⇒ {I10} ⇒ {R63}	{Overweight and obesity} ⇒ {Essential (primary) hypertension} ⇒ {Symptoms and signs concerning food and fluid intake}
		{E66} ⇒ {E78} ⇒ {R63}	{Overweight and obesity} ⇒ {Disorders of lipoprotein metabolism and other lipidemias} ⇒ {Symptoms and signs concerning food and fluid intake}
Adolescence	2019/20	{M26} ⇒ {R63}	{Dentofacial anomalies [including malocclusion]} ⇒ {Symptoms and signs concerning food and fluid intake}
	2020	{U07} ⇒ {Z88} ⇒ {Z20}	{Covid-19} ⇒ {Allergy status to drugs, medicaments and biological substances} ⇒ {Contact with and (suspected) exposure to communicable diseases}
		{R05} ⇒ {U07} ⇒ {Z20}	{Cough} ⇒ {Covid-19} ⇒ {Contact with and (suspected) exposure to communicable diseases}
Adulthood	2020	{R05} ⇒ {R50} ⇔ [{I65} ⇔ {J40} ⇔ {R50}]	{Cough} ⇒ {Fever of other and unknown origin} ⇔ [{Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction} ⇔ {Bronchitis, not specified as acute or chronic} ⇔ {Fever of other and unknown origin}]
		{Z57} ⇒ {Z77} ⇔ [{J00} ⇔ {Z77} ⇔ {D18}]	{Occupational exposure to risk factors} ⇒ {Other contact with and (suspected) exposures hazardous to health} ⇔ [{Acute nasopharyngitis [common cold]}⇔{Other contact with and (suspected) exposures hazardous to health}⇔{Hemangioma and lymphangioma, any site}]
Elderly	2019/20	{E87} ⇒ {I10} ⇒ {I52}	{Other disorders of fluid, electrolyte and acid-base balance} ⇒ {Essential (primary) hypertension} ⇒ {Other heart disorders in diseases classified elsewhere}
		{K52} ⇒ {R19}	{Other and unspecified noninfective gastroenteritis and colitis} ⇒ {Other symptoms and signs involving the digestive system and abdomen}

Table 7.7: **Directionality paths pre/post COVID-19 outbreak.** Some interesting paths found in AGES datasets. AGES yearly comparison directionality network.

7.5 Discussion

The procedure correctly states an initial basic formulation of the directionality detection to improve the visualization and interpretation of association rule networks. The goal is based in the intuition that while rules detect non-temporal associations, by detecting the directionality, we are able to detect not only temporal association, but also possible comorbidity diagnoses. The intuition needs further research to incorporate mathematical and statistical propositions in the fields of bayes networks, process mining, sequence mining and causality detection to improve the initial formulation. These mathematical scope is out of scope of the initial formulation proposed in this paper, which only provides with an initial suggestion of diectionality.

In most of the cases, the algorithm has proven correctness, untying the associations and providing a clearer network with respect the rule network. Not only providing association, but also suggesting directionality/temporality or comorbidity in the case where there is not a clear order of appearance.

Regarding detection in year 2020 within covid-19 context, the algorithm has not meet the expectations. Due to the fact that the trajectories are formed using data from the active diagnoses, we are expanding longer in time than the both years studied, with the trajectories not changing much from 2019 to 2020 (since a patient with a subset of active diagnoses would present almost the same trajectory in both years). The algorithm only has found some rules related to covid codes but has correctly identified changes in the associations of other diagnoses due to the shits in medical priorities suffered in the outbreak.

Chapter 8

Medical evaluation

The following chapter is devoted to give a short review of the methodology addressed from a medical perspective. All the previous chapters have been consciously evaluated by medical experts to assess the correctness of the resulting rules. Some of the strengths and weaknesses found from a medical point of view are outlined in this section.

Medical experts found interesting the approach being tackled in this study. Although the process might not be a claim by the medical community practitioners found interesting to have these utilities to be able to further explore possible medical issues.

The medical community could develop synergies with the data-science community to provide the healthcare system with further robust techniques to improve already existing procedures or explore new possibilities in the field, since account with high volumes of data that need to be mined. Moreover, the medical knowledge is needed to refine these designed approaches

The appearance of these techniques could be a good starting point to raise awareness about the importance of the coding. With the adoption of new technological procedures in the healthcare system, some practitioners do not give the importance needed to the coding process, sometimes not coding a visit or incorrectly using some codes to relate different conditions. This will always be the last limitation of those models since ultimately rely on the assumption that the practitioner is coding correctly each condition. The implementation of those techniques may not be useful for all practitioners, but could set a precedent to raise the aforementioned awareness for coding.

8.1 Association rules as unsupervised learning methods

It is important to keep in mind that this thesis is only intended to provide an initial formulation of a technique, further analysis must be done to achieve a point where the medical community could take advantage of the approach. Therefore, although the ultimate goal designing a methodology used for unsupervised learning, by imposing highly demanding thresholds we alter the behaviour of it to turn it into a "supervised learning method", aiming to find obvious rules that could validate the correctness of the model. In posterior steps, the thresholds should be softened to mate the same objectives as those for which it was created.

The validation procedures and results have proven correctness of the models, with almost all rules found known or intuited beforehand.

The networks obtained from adult and elderly datasets are specifically accurate, presenting a clearer design and interactions for codes relating important chronic conditions.

In the case of children and adolescents the results are less suggestive, possibly requiring higher thresholds. There is a consensus from the experts in mentioning that these segments of the population tend to encounter the healthcare systems for different conditions that might happen by change, therefore the visits do not present a pattern of behaviour. Moreover, they are visited lesser than higher age groups and due to the age present shorter trajectories with non or very few active diagnoses (which turn to be the ones interesting, coding for important conditions spanning long in time).

The visualization framework is attractive and easy to interpret by the practitioners.

8.2 Morbidity suggestion

There is agreement on the interestingness of the multimorbidity approach. Nevertheless the experts point that the algorithm relate similar conditions rather than multimorbidity groups. Several "in silico" an "in vivo" studies must be carry on in parallel to inferr a multimorbidity group. NEvertheless it is interesting to discretize the network in groups for better understanding the groups of conditions associated.

8.3 Directionality suggestion

There is consensus on the feel that the directionality can improve the visualization, adding further knowledge to the networks.

Some directionalities have created polemic, turning to suggesting wrong the temporality. This is produced by the initial though that *active diagnoses* of each visit where stored in order of appearance. After addressing the problem with the IT department, it is noted that the *active diagnoses* follow an importance order based in an index internally stored at ICS, thus making impossible to detect the correct directionality. There is an agreement on the fact that the trajectories must be reformulated, using information from more years to loose minimal information and catch the correct order of appearance. Moreover, the definition of an scale of code-importance for population segment should ease the removal of those uninteresting codes that create noise in the case of children and adolescence for example.

Those directionalities relating codes separated long in time must not be directed, since they are probably not related.

The practitioners found the procedure interesting but agree that more information for each directionality should be retrieved by the model to improve the knowledge. The computation of the lift, support and confidence should mate these expectations.

8.4 Covid evaluation

In the case of the covid evaluation, the result is not the one wanted. When exploring small periods of time, one should only take into account the diagnoses coded in those periods to be able to detect interesting associations. If to the contrary, we explore the trajectories of the patients of 2019 and 2020, adding the active diagnoses, these does not change enough for the algorithm to detect associations of covid-19 period. A new formulation of the trajectories is needed to adress these research questions.

Chapter 9

Conclusion

This chapter is devoted to outline the principal conclusions obtained after the realization of the master thesis

Rule miners, although being powerful non-supervised statistical methods, are not among the principal techniques used by data scientists. Traditionally, the well-known problems of the parametrization expertise, the dimension of the output and the difficulty to find a dataset matching the form of the input knocks back some off the experts when it comes to carrying out a mining. Nevertheless, the proposed approach used in this thesis solves the major drawbacks stated before, providing a better, most robust technique to mine datasets as the one used in this thesis, with very little expertise of the technique.

The initial objective of this thesis was twofold:

First, testing the Yacaree framework designed by professor José Luis Balcázar, conducting yet another evaluation of the methodology in a real case dataset to prove its validity and find weaknesses. The ultimate goal was to demonstrate the algorithm is able to tackle the principal limitations aforementioned by rule miners, by automatization of the principal threshold and setting up a redundancy notion which is able to reduce the amount of output rules without loosing much information.

Second, relying on an internship in the ICS, another goal was to set a basis of a technique usage not done before in the area of interest to mine a medical dataset to be able to identify associations of diagnoses (more specifically, chronic diseases). The final goal is to create a useful technique, with an easy interpretation for decision making in the medical Catalan healthcare system.

The initial objectives have been achieved, demonstrating the technique is reliable, robust and valid to detect associations that can be useful by the medical community. Nevertheless, there is still road ahead.

Yacaree has proven validity at detecting a human redeable amount of medical associations that turn to be correct based on medical knowledge. Moreover, when comparing with other classical rule miners such as "apriori", yacaree provides less rules irredundant between them that are significant based on the thresholds. The effect of the confidence boost is clear from the context. High demanding confidence boost thresholds tend to reach a point of maximum irredundancy where the amount of rules is not lowered even if increasing the thresholds. The results suggest that using this framework and lowering the confidence thresholds, would provide some interesting rules that are not evident or known before hand by the community, while the possibility to increase the confidence boost will reduce this amount to an irredundant base, but still providing the interesting associations.

Yacaree cannot compete with other rule miners in terms of efficiency, lasting much more. While this is not a problem for the adressed dataset, it can discard the procedure for other experimentations depending on the context.

The algorithm works well at detecting between and within associations of diagnoses in groups of population. In one hand, if the algorithm is used in a dataset comprising the medical trajectories of all the population, the result will cover associations between diagnoses and groups of population, or very important associations of conditions found in all segments of population. In the other hand, if the algorithm is used in a controlled group of population (elderly, male, etc.) the result will cover associations of diagnoses within the group. In all the cases, the associations retrieved turned to be correct medically speaking. This behaviour suggests the algorithm could perform well at detecting associations with a desired diagnostic if controlling for population experiencing it, avoiding the effect of the centers of gravity described in 5.1.3.

By using network visualization techniques we are able to interpret a higher number of rules from a macro point of view, providing an attractive methodology that ease the interpretation for practitioners. The algorithm retrieves better results for groups of populations that are better defined and medically contextualized. With the aging, patients present longer trajectories with more cooccurring conditions. The algorithm provides more significant rules in this cases, which are more interesting and better known by the medical community. Results suggest the thresholds must be raised in more diffuse groups such as childhood or adolescence.

A suggestion of multimorbidity detection based on hierarchical clustering is proposed. The algorithm correctly provides groups of related diagnoses. Although the results are promising, much more researh must be done and we cannot state the algorithm identifies morbidity, since is a field that needs a lot of research to detect groups of cooccurring diagnoses.

A proposal of a method to detect directionality is suggested. From a medical point of view, it is not only important to detect co-occurrence or association between diagnoses, but to also detect the temporality or order of appearance of the algorithm. Therefore, the goal is to detect those associations of conditions and suggest an order of appearance to increase the knowledge and improve the visualization. The algorithm appear to detect correctly directionality based on probability but more work must be done to include notions from bayesian networks, sequence mining and process mining that are out of scope for this thesis. Moreover, trajectory construction must be reevaluated since a part of the diagnoses is not temporarily ordered, thus failing at detecting directionality.

Finally, the method is evaluated to check wheter the algorithm is able to detect a change of associations in the context of covid. The algorithm does not work well since the trajectories gather diagnoses spanning several years. Therefore the diagnoses from the covid year only comprise a small part of the trajectories. Some of the associations are related to covid but to achieve better results the method should be tested on partial trajectories for each year. However, the procedure is able to add more information to the networks by adding yearly data to the associations.

Chapter 10

Future work

Due to the limitations of resources and time, this thesis only attain a first approximation to the model and the procedures. Moreover, more testings and efforts will be needed in further steps to achieve a useful proposition for the medical commbvhunity. This section is devoted to propose some steps to be a continuation of this work, that could benefit the procedure.

10.1 Algorithm

- An improvement of the suggested *directionality algorithm* in section 7 should make it more robust and efficient. It is needed more research in the topics of *bayesian networks* and *process mining* to improve the flow charts and inferr correctly the causality. A simple model using binomial tests turns to be effective for this thesis since it is not the initial objective of the research, but in order to create a tool that can be useful and fully correct for the practitioners, more research and formulations must be done to make it more robust.

- *Analyze the parametrization.* The algorithm is tested with high demanding thresholds in this thesis. These thresholds are intentionally set to provide obvious rules. This will demonstrate the validity of the model, although not providing rules interesting in terms of research, not seen at a glance. If the algorithm works well at detecting obvious rules one could extrapolate the algorithm will found non-obvious rules when softening the thresholds. Some intuitions are that the confidence of the rules could be softened to around 60% or 65% provided that the boost increases considerably (say to 1.50 or more). This will increase the number of rules but avoid a high portion of redundant ones. Detecting lower non-redundant associations.

- The *multimorbidity detection proposition* in section 6.2.2, although valid, is based in a preliminary intuition. Further research must be done to detect multimorbidity groups. Moreover, the same procedure has to be applied to the directionality plots, to check the validity. Same, we should label the edges with the support in order to provide a weight to it, to improve the multimorbidity detection, by setting lengths to the edges between diagnoses.

- Add different parameters to the associations in the model. Add support, confidence, lift and other interesting parameters to the associations for pruning, prediction and knowledge discovery.

10.2 Testing confounding factors

- The confounding factors *sex and age* are tested separately. However, both factors are highly related. A good approach will be to test the combinations of both factors. Some intuitions suggest that in the case of the $ADULT$ population, there are high differences between $MALES$ and $FEMALES$. For example in the case of *pregnancy*. Several rules relating diagnoses of pregnancy are avoided when don't taking into account the confounding factor AGE. Other intuitions are the *alimentary disorders* in the case of $ADOLESCENT\ FEMALES$, that are more prevalent in this sex than in the other.

- The algorithm has shown promising thresholds when partitioning the datasets based on confounding factors *sex and age*, by don't masking the associations related to a certain group. Some intuitions suggest that the same can be made to certain diseases, for example *chronic diseases*. By mining all the *"medical trajectories"* in the dataset presenting certain chronic condition, one is able to detect the associations suffering the population with the condition.

10.3 Dataset

- Create a more clear, more robust definition of the medical trajectories. There are some conditions that must not being taken into account when constructing the trajectories. Some visit diagnoses are not relevant in the medical life of a patient. While a "fall" is not relevant for a child, because most of the time will not have a relevant effect in the life of the patient, in the case of elderly it could lead to other conditions possible inducing health hazards. Therefore, there is needed a further exploration of the codes of diagnoses, proposing a grading on the "importance" or "chronicity" or "relevancy" of each diagnoses, based on confounding factors. This could also be useful for the medical health services.

- Add a temporality proposition to the medical trajectories. It is not important the associations between codes that are separated a lot in time in most of the cases. Create a variable measuring temporality.

- It is needed better datasets. In one hand, practitioner need to see more tools like this to raise awareness of the importance of codification at all levels, not only in the leading hospitals but also in the primary health services. If practitioners codify correctly, these data is a gold resource for mining and return knowledge. The dataset used, counts for a lot of miss codification, some diagnoses act as a hotpocht such as the ones detected in section 5.1.3. practitioner need more time to detect conditions in the patients and correctly diagnose those visits. The testing should be done using also the dataset from these leading hospitals to improve the knowledge.

- After studying the database and talking with the quality managers of the database it is pointed that while visit diagnoses follow a chronological order, the active diagnoses are stored in the database based on an importance index. Therefore it is needed to use more data to construct the medical trajectories in the correct order. Some of the associations, the ones regarding diagnoses extracted from the visits follow a correct order, thus the directionality is correct. Nevertheless, for those diagnoses extracted from the active diagnoses of the patients, lacking teporality order, the directionality plots suffer from the incorrectness of the data.

- When comparing 2019 with 2020, one should only use the visit diagnoses, since we are trying to evaluate infradiagnosing of associations. If we use active diagnoses we may be using the same information in both years, since a patient with the same chronic diseases might have the same trajectory in both years.

10.4 Research

- Evaluate the performance when only exploring diagnoses. Not taking into account socioeconomic factors nor sociodemographic factors. It must be noted that when using these factors with the diagnoses data, we are not only detecting morbidity, we are also detecting *patient complexity*. Most effort has to be done to get rid of those codes that are not diagnoses, mainly Z codes.

- Evaluate the performance when adding more *social, environmental and behavioral factors*. One example would be to use data about the GDP, smoking habits... This data is available at ICS dataset and could be used to gain knowledge about morbidities in these groups. Can a drug from one condition lead to another??.

- It will be interesting to use also *pharmacological data* to gain knowledge about *causality or co-occurence* between drugs and conditions. Some intuitions suggest this could be good research line.

- Some intuitions also suggest that adding *procedimental data* could help to make a map of the conditions and the procedures being held in the medical services. This could act as a map of the medical procedures, gathering the costs and probabilities of paths.

10.5 Medical evaluation

- Use medical evaluation to improve the whole procedure and guide the research.

10.6 Visualization

- Improve visualization techniques. Add comorbidity groups to the associations, use colors to give information about lift/support/confidence,... Check other visualization tools in other computational languages. Create a shiny application gathering all the networks.

Bibliography

[1] Agnieszka O, Marek H. (1999). *A Bayesian network model for diagnoses of liver disorders.*.

[2] Agrawal R, Imieliński T, Swami A. (1993). *Mining association rules between sets of items in large databases*. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p.207. doi:10.1145/170035.170072.

[3] Asuncion A, Newman D. (2007). UCI machine learning repository.

[4] Balcázar JL. (2013). *Formal and computational properties of the confidence boost of association rules.* ACM Trans. Knowl. Discov. Data 7, 4, Article 19 (November 2013), 41 pages. DOI:https://doi.org/10.1145/2541268.2541272

[5] Balcázar JL. (2009). *Two measures of objective novelty in association rule mining.* In PAKDD Workshops (Springer-Verlag LNCS 5669). 76-98.

[6] Balcázar JL. (2009). *Two measures of objective novelty in association rule mining.* In PAKDD Workshops (Springer-Verlag LNCS 5669). 76-98.

[7] Balcázar JL. (2010a). *Closure-based confidence boost in association rules. JMLR Workshop and Conference Proceedings* Workshop on Applications of Pattern Analysis 11, 1-7.

[8] Balcázar JL. (2010b). *Objective novelty of association rules: Measuring the confidence boost.* In EGC, S. B. Yahia and J.-M. Petit, Eds. Revue des Nouvelles Technologies de l'Information Series, vol. RNTIE-19. Cépaduès-Éditions, 297-302.

[9] Balcázar JL. (2010c). *Redundancy, deduction schemes, and minimum-size bases for association rules.* Logical Methods in Computer Science 6, 2:3, 1-33.

[10] Balcázar JL. (2011). *Parameter-free association rule mining with yacaree.* See Khenchaf and Poncelet [2011], 251-253.

[11] Balcázar JL, Tîrnˇaucˇa C. (2011). *Closed-set-based discovery of representative association rules revisited.* See Khenchaf and Poncelet [2011], 635-646.

[12] Balcázar JL, Tîrnˇaucˇa C, Zorrilla M. (2010a). *Mining educational data for patterns with negations and high confidence boost.* Taller de Minería de Datos TAMIDA 2010; available at: [http://personales.unican.es/balcazarjl].

[13] Balcázar JL, Tîrnˇaucˇa C, Zorrilla M. (2010b). *Filtering association rules with negations on the basis of their confidence boost.* KDIR 2010. Available at: [http://personales.unican.es/balcazarjl].

[14] Bayardo R, Agrawal R, Gunopulos D. (1999). *Constraint-based rule mining in large, dense databases.* In ICDE. 188–197.

[15] Ben Gal I. (2007). *Bayesian Networks.* In Ruggeri F, Kennett RS, Faltin FW (eds.). Support-Page. Encyclopedia of Statistics in Quality and Reliability. John Wiley Sons. doi:10.1002/9780470061572.eqr089. ISBN 978-0-470-01861-3.

[16] Bozkaya M, Gabriels J, van der Werf JM. (2009). *Process diagnoses: A Method Based on Process Mining.* 2009 International Conference on Information, Process, and Knowledge Management, 2009, pp. 22-27, doi: 10.1109/eKNOW.2009.29.

[17] Brossette S, Sprague A, Hardin J, Waites K, Jones W, Moser S. (1998). *Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance.* Journal of the American Medical Informatics Association : JAMIA. 5. 373-81. 10.1136/jamia.1998.0050373.

[18] Capobianco E, Lio P. (2013). *Comorbidity: a multidimensional approach.* Trends in Molecular Medicine, Volume 19, Issue 9, 2013, Pages 515-521, ISSN 1471-4914, https://doi.org/10.1016/j.molmed.2013.07.004.

[19] Diederichs C, Berger K, Bartels DB. (2011). *The Measurement of Multiple Chronic Diseases—A Systematic Review on Existing Multimorbidity Indices.* The Journals of Gerontology: Series A, Volume 66A, Issue 3, March 2011, Pages 301–311, https://doi.org/10.1093/gerona/glq208

[20] Groll DL, To T, Bombardier C, Wright JG. (2005). *The development of a comorbidity index with physical function as the outcome.* Journal of Clinical Epidemiology, Volume 58, Issue 6, 2005, Pages 595-602, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2004.10.018.

[21] Held F, Blyth F, Gnjidic D, Hirani V, Naganathan V, Waite L. (2016). *Association Rules Analysis of Comorbidity and Multimorbidity: The Concord Health and Aging in Men Project.* The Journals of Gerontology: Series A, Volume 71, Issue 5, May 2016, Pages 625–631

[22] Kryszkiewicz M. (2001). *Closed set based discovery of representative association rules.* In Proc. of the 4th International Symposium on Intelligent Data Analysis (IDA), F. Hoffmann, D. J. Hand, N. M. Adams, D. H. Fisher, and G. Guimaraes, Eds. Lecture Notes in Computer Science Series, vol. 2189. Springer-Verlag, 350–359.

[23] Kumar Y, Sahoo G. (2013). *Prediction of different types of liver diseases using rule based classification model.* Technol Health Care, 21 (5) (2013), pp. 417-432.

[24] Lakshmi KS, Vadivu G. (2009). *A novel approach for disease comorbidity prediction using weighted association rule mining.* J Ambient Intell Human Comput (2019). https://doi.org/10.1007/s12652-019-01217-1

[25] Lakshmi KS, Vadivu G. (2017). *Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis.* Procedia Computer Science, Volume 115, 2017, Pages 290-295, ISSN 1877-0509,

[26] Lefèvre T, d'Ivernois JF, De Andrade V, Crozet C, Lombrail P, Gagnayre R. (2014). *What do we mean by multimorbidity? An analysis of the literature on multimorbidity measures, associated factors, and impact on health services organization.* Revue d'Épidémiologie et de Santé Publique, Volume 62, Issue 5, 2014, Pages 305-314, ISSN 0398-7620, https://doi.org/10.1016/j.respe.2014.09.002.

[27] Liu B, Hsu W, Ma Y. (1999). *Pruning and summarizing the discovered associations.* In Proc. Knowledge Discovery in Databases. 125–134.

[28] Lopez Segui F, Hernandez Guillamet G, Pifarré Arolas H, Marin Gomez X, Ruiz Comellas A, Ramirez Morros AM, Adroher Mas C, Vidal-Alaball J. (2021). *Big data-based analysis to characterise and identify variations in the type of Primary Care visits before and during COVID in Catalonia.*Journal of Medical Internet Research. 12/06/2021:29622 (forthcoming/in press)

[29] Luxenburger M. (1991). *Implications partielles dans un contexte.* Mathématiques et Sciences Humaines 29, 35–55.

[30] Mazarbhuiya F, Alzahrani M. (2020). *Breast Cancer diagnoses using Sequential Pattern Mining.* International Journal of Recent Technology and Engineering. 8. 10.35940/ijrte.F9171.038620.

[31] Mclachlan S, Graham F, Norman E. (2020). *Bayesian Networks in Healthcare: Distribution by Medical Condition.* .

[32] Padmanabhan B, and Tuzhilin A, (2000). *Small is beautiful: discovering the minimal set of unexpected patterns*. In Proc. Knowledge Discovery in Databases. 54–63.

[33] Pearl J. (2000). *Causality: Models, Reasoning, and Inference.*. Cambridge University Press. ISBN 978-0-521-77362-1. OCLC 42291253.

[34] Pifarré i Arolas H, Vidal-Alaball J, Gil J, López F, Nicodemo C, Saez M. (2021). *Missing Diagnoses during the COVID-19 Pandemic: A Year in Review*. International Journal of Environmental Research and Public Health. 2021; 18(10):5335. https://doi.org/10.3390/ijerph18105335

[35] Pombo N, Garcia N, Bousson K. (2017). *Classification techniques on computerized systems to predict and/or to detect Apnea: A systematic review*. Comput Methods Programs Biomed, 140 (2017), pp. 265-274.

[36] Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. (2014). *Multimorbidity patterns: a systematic review*. Journal of Clinical Epidemiology, Volume 67, Issue 3, 2014, Pages 254-266, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2013.09.021.

[37] Rashidi P. (2014). *Chapter 5 - Stream Sequence Mining for Human Activity Discovery*. Plan, Activity, and Intent Recognition, Morgan Kaufmann, 2014, Pages 123-148, ISBN 9780123985323, https://doi.org/10.1016/B978-0-12-398532-3.00005-1.

[38] Schäfer I, von Leitner EC, Schön G, Koller D, Hansen H, Kolonko T, Kaduszkiewicz H, Wegscheider K, Glaeske G, van den Bussche H. (2010). *Multimorbidity Patterns in the Elderly: A New Approach of Disease Clustering Identifies Complex Interrelations between Chronic Conditions* . PLOS ONE. December 29, 2010 https://doi.org/10.1371/journal.pone.0015941

[39] Shah D, Lakshmanan L, Ramamritham K, Sudarshan S. (1999). *Interestingness and pruning of mined patterns.* In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

[40] Shen Y, Zhang L, Zhang J, Yang M, Tang B, Li Y, Lei K. (2018) *CBN: Constructing a clinical Bayesian network based on data from the electronic medical re-cord.*

[41] Spirtes P, Glymour CN, Scheines R (1993). *Causation, Prediction, and Search (1st ed.)*. Springer-Verlag. ISBN 978-0-387-97979-3.

[42] Srinivas D, Ravi A, Torney D. (2001). *Discovery of Association Rules in Medical Data. Medical informatics and the Internet in medicine.* 26. 25-33. doi:10.1080/14639230010028786.

[43] Stilou S, Panagiotis M, Pappas C. (2001). *Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare.* Studies in health technology and informatics. 84. 1399-403.

[44] Tai YM, Chiu HW. (2009). *Comorbidity study of ADHD: Applying association rule mining (ARM) to National Health Insurance Database of Taiwan.* International Journal of Medical Informatics, Volume 78, Issue 12, 2009. Pages e75-e83, ISSN 1386-5056.

[45] Toivonen H, Klemettinen M, Ronkainen P, Hatôonen K, Mannila H. (1995). *Pruning and grouping discovered association rules.* In ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases. 47–52.

[46] Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. (2009). *Defining comorbidity: implications for understanding health and health services.* Ann Fam Med. 2009;7(4):357-363. doi:10.1370/afm.983

[47] van den Bussche H, Koller D, Kolonko T. (2011). *Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? Results of a claims data based cross-sectional study in Germany.* BMC Public Health 11, 101 (2011). https://doi.org/10.1186/1471-2458-11-101

[48] van der Aalst W. (2011). *Process Mining: Data Science in Action.*

[49] van der Aalst W. (2016). *Process Mining: Data Science in Action.*

[50] Verma T, Pearl J (1991). *Equivalence and synthesis of causal models.* In Bonissone P, Henrion M, Kanal LN, Lemmer JF (eds.). UAI '90 Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. Elsevier. pp. 255–270. ISBN 0-444-89264-8.

[51] Wang CH, Lee TY, Hui KC, Chung MH. (2019). *Mental disorders and medical comorbidities: Association rule mining approach.* Perspect Psychiatr Care. 2019 Jul;55(3):517-526. doi: 10.1111/ppc.12362. Epub 2019 Feb 7. PMID: 30734309.

[52] Zaki MJ. (2004). *Mining non-redundant association rules.* Data Min. Knowl. Discov. 9, 3, 223–248.

[53] Zamora Casals M, Gavaldà Mestre R. (2017). *Computing and visualizing informative trajectories in temporaly annotated data.* UPCCommons. URIhttp://hdl.handle.net/2117/109799

[54] *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. `http://www.cdc.gov/nchs/icd/icd10cm.htm`