Master in Innovation and Research in Informatics
Data Science

Acquisition of Patterns from Medical Records

Master Thesis

Author:

Oriol Borrell Roig
FIB - UPC Student
Barcelona, Spain
`oriol.borrell.roig@estudiantat.upc.edu`

Supervisors:

Alicia Ageno Pulido
Computer Science Department
Barcelona, Spain
`ageno@cs.upc.edu`

Neus Català Roig
Computer Science Department
Vilanova i la Geltrú, Spain
`ncatala@cs.upc.edu`

June 28, 2021

Facultat d'Informàtica de Barcelona (FIB)
Universitat Politècnica de Catalunya - BarcelonaTech
(UPC)

Acknowledgments

First of all, I would like to express my gratitude to my thesis supervisors: Alicia Ageno Pulido and Neus Català Roig from the UPC. I want to thank them for their patient guidance, supervision, and advice throughout the development of this master's thesis. It was my privilege to work with them. I want also to thank professor Jordi Turmo for the support received during the evaluation of the results.

Second, to IDIAP (Institut Universitari d'Investigació en Atenció Primària) for sharing the data and other resources used in this project. I want to do a special menton to Pere Toran, Concepció Violan, Víctor Miguel López, David Lacasta, and Luis Rodriguez, doctors of the IDIAP Jordi Gol Institute, for the support received during the evaluation of the results.

Third, to all the professors that provided me the necessary knowledge to overcome the difficulties encountered during this Data Science master's degree. Also, to my classmates that made this experience much more enriching from a personal point of view. I wish them all the best in life.

Forth, to my friends especially to Marina Prieto and Beatriz Fernandez-Montells, for gathering me some medical knowledge that I did not have.

And finally, to my family, for supporting and providing me the best education they could afford.

Abstract

In recent years, the volume of information available electronically has increased exponentially, and the field of primary health care has not been an exception. The increasing availability of this electronic data, represents an impact on the potential discovery of patterns to predict the risk of new diseases, helping the personalized care and increasing the quality of life. Extracting frequent patterns from medical records represents a huge challenge in Data Mining, knowing that in this context the analysis of the temporality between clinical instances is a must.

In the TADIA-MED research project, data containing information on visits of patients at Primary Care Centers (CAP) throughout Catalonia was obtained. All annotations in the textbook that the doctor registers in the health system during visits follow what is called the MEAP structure (**M**otiu de la consulta, **E**xploració, **A**valuació i **P**la d'actuació, in Catalan). The information contained in these MEAPs was classified into *Diagnostics*, *Signs or symptoms*, *Drugs*, or *Body parts*. This information was represented as a graph and stored in a Neo4J server.

In this thesis, a new formulation is presented which defines how to compute the temporal association rules in the explained context. The obtained rules are intended to be diagnostic aid patterns. We also have developed an algorithm that uses our formulation to extract the temporal rules. This algorithm makes it possible to parameterize the desired rules in various aspects with respect to the desired format or temporality. We are also capable of extracting rules at different levels of abstraction. Finally, we have defined a process for evaluating the rules obtained. The designed process will be the evaluation process of the entire TADIA-MED project. In spite of the small volume of available data, the evaluation of the rules obtained has been very promising and will help us to continue improving.

Contents

1 Introduction 6
1.1 Context of the Project . . . 6
1.2 Goals . . . 7
1.3 Previous Note . . . 7

2 State of the art 8
2.1 Apriori . . . 9
2.2 Handling different type of temporal data . . . 10
2.3 Temporal associations rules with multiple antecedents . . . 11

3 Methodology 14
3.1 Strategy . . . 14
3.2 Technologies . . . 15
3.2.1 MyDisk . . . 15
3.2.2 Neo4J . . . 15
3.2.3 Python . . . 15

4 Development 16
4.1 Available data . . . 16
4.1.1 Node properties . . . 17
4.1.2 Interaction between nodes . . . 22
4.2 Data extraction . . . 23
4.3 First exploratory data analysis . . . 24
4.4 Data preparation . . . 27
4.4.1 Preprocessing strings . . . 27
4.4.2 Generalization of codes . . . 27
4.5 Second exploratory Data Analysis . . . 30
4.6 Our temporal association rules formulation . . . 33
4.7 Temporal association rules algorithm . . . 34
4.7.1 Principal data structures . . . 34
4.7.2 Temporal association rules extraction . . . 35

5 Evaluation 38
5.1 Description of the evaluation phase . . . 38
5.2 Results . . . 40

6 Conclusions 45
6.1 Future Work . . . 46

References 48

A Time intervals and relevance of diagnoses 50

B Composition Pack 1 **61**

C Composition Pack 2 **66**

D Evaluations Pack 1 **71**

E Evaluations Pack 2 **76**

List of Figures

1 Data model 17
2 ATC7 generalization tree summary 19
3 CIAP2 generalization tree summary 20
4 CIM10 generalization tree summary 22
5 Instances from patient 32, days 20161109, 20161110, 20161111 and 20161112 23
6 CSV file obtained after extracting the data 24

List of Tables

1 Available data for Farmaco 18
2 Available data for SignoSintoma 19
3 Available data for ParteCuerpo 20
4 Available data for Diagnostico 21
5 Episodes and Variables distribution before preprocessing 24
6 10 most frequent values for Farmaco nodes 25
7 10 most frequent values for SignoSintoma nodes 25
8 10 most frequent values for ParteCuerpo nodes 26
9 10 most frequent values for Diagnostico nodes 26
10 Errors found when generalizing CIM10 code 28
11 Errors found when generalizing CIAP2 code 29
12 Episodes and Variables distribution after preprocessing 30
13 10 most frequent values for Farmaco nodes after preprocessing 31
14 10 most frequent values for Diagnostico nodes after preprocessing 31
15 10 most frequent values for SignoSintoma nodes after preprocessing . . . 32
16 10 most frequent values for ParteCuerpo nodes after preprocessing 32
17 Original data 34
18 Temporal data structure 34
19 Binary data structure 35
21 Summary of the results, pack 1 41
23 Agreement between doctors in pack 1 42
25 Summary of the results, pack 2 43
27 Similarities between doctors in pack 2 43
28 Recommended time gap and importance per diagnosis 60
29 First pack results 65
30 Second pack results 70
31 Evaluation of the first pack results 75
32 Evaluation of the second pack results 81

1 Introduction

Over the last few years, there has been an increasing attention in the collection of data from healthcare institutions. Given that the healthcare delivery process is not characterized by a "static" approach but evolves over time, capturing the temporal relationships between interesting events is a crucial point for forecasting and planning operational policies.

The application of Data Mining (DM) techniques can be particularly suitable for the extraction of meaningful information and knowledge from large healthcare databases. In the literature, there are several forms of definitions for temporal association rules which have different formulas for calculating support and confidence. The exploitation of Temporal Data Mining (TDM) methods able to deal with time and temporal events could be particularly fruitful in this field.

In this project, we aim to explore approaches that allow the extraction of patient evolution patterns from clinical histories written in Spanish, Catalan, or English. After reviewing the literature, we will propose our formulation of temporal association rule which was equipped with a new formula for calculating its support and confidence.

Our method will be capable of parametrizing the time of the rules and extract useful time information. We will also be able to obtain rules using the generalization of events in a higher level of abstraction. The obtained patterns could be useful for the development of diagnostic assistants or prevention policies.

1.1 Context of the Project

The present document was carried out during the second semester of course 2020/21 and presented as the master thesis of the Master in Innovation and Research in Informatics. We must understand this thesis in the context of the TADIA-MED [17] project, developed by the Language and Speech Technologies and Applications (TALP) [18] research center at the Universitat Politècnica de Catalunya (UPC).

The TADIA-MED [17] project has the aim to explore approaches that allow the extraction of patient evolution patterns from clinical histories. The TADIA-MED project is focused on the study of different aspects:

- Medical information extraction from clinical histories.
- Negation and speculation detection.
- Enrichment and approximated term search in medical ontologies.
- Knowledge discovery for risk prediction in multimorbid patients.

We can understand this thesis in the last focus of study: *"Knowledge discovery for risk prediction in multimorbid patients"*. Multimorbid patients are highly prevalent in some clinical contexts, such as primary care, but there is little evidence about how to deal with such patients. In collaboration with IDIAP Jordi Gol [9], we aim at automatically inferring patterns by which doctors can predict the risk of new diseases for a multimorbid patient given her clinical history.

1.2 Goals

The main project goal is to find a robust formulation that lets us obtain temporal association rules in multi-attribute data. We will make a special effort in the time factor: in being able to parameterize it, and in having a notion of the time elapsed in the results. We also want to have the possibility to express the different medical instances in different levels of abstraction, allowing the generalization of instances. We will use our framework to acquire patterns to aid in diagnosing.

Beforehand, we will perform a state-of-the-art study of temporal association rules mining. Next, we will perform an exploratory data analysis process to study our dataset to define our association rule framework. Finally, we will apply our formulation to our data.

To measure the correctness of our framework, the results will be analyzed by doctors of the IDIAP Jordi Gol [9] center.

These goals are going to be detailed and explained in the following sections.

1.3 Previous Note

This document contains some medical concepts explained in different languages: Catalan, Spanish, and English. I will not translate them for two reasons. The first one is because we understand our corpus is composed of those terms, and if we modify them we would not be showing our corpus. If despite the first reason we wanted to translate them, they are very concrete terms, and I don't have the required medical knowledge to perfectly translate those concepts.

2 State of the art

Association rule mining, raised by Rakesh Agrawal, is an important research problem in the data mining field. Association rule mining aims at detecting the relationship of tuples in a transactional database and serving decision making [1].

The author defined the problem as follows: Let $I = i_1, i_2, ..., i_m$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. We say that a transaction T *contains* X, a set of some items in I, if $X \subseteq T$. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = 0$. X is called the *antecedent* or *left-hand-side (LHS)* of the rule and Y is called the *consequent* or *right-hand-side (RHS)* of the rule.

For each association, several indicators can be computed:

- *Support of an itemset:* The support value of X with respect to T is defined as the proportion of transactions (instances) in the database, which contains the itemset X with respect to the total number of instances.

$$support(X) = \frac{|Transactions\ containing\ X|}{|T|} \tag{1}$$

- *Support of a rule:* The support value of a rule, $X \Rightarrow Y$, with respect to T is defined as the percentage of all transactions in the database, which contains the itemset X and the itemset Y.

$$support(X \Rightarrow Y) = P(X \cup Y) = \frac{support(X \cup Y)}{|T|} \tag{2}$$

- *Coverage of a rule:* Coverage is sometimes called antecedent support or LHS support. It measures how often a rule, $X \Rightarrow Y$, is applicable in a database.

$$coverage(X \Rightarrow Y) = support(X) \tag{3}$$

- *Confidence of a rule:* The confidence value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T, is the proportion of the transactions that contain X, which also contain Y.

$$confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{support(X \cup Y)}{support(X)} \tag{4}$$

- *Leverage of a rule:* The Leverage value of a rule, $X \Rightarrow Y$, measures the difference between the probability of the rule and the expected probability if the items were statistically independent. It ranges from $[-1, +1]$ indicating 0 the independence condition.

$$leverage(X \Rightarrow Y) = support(X \Rightarrow Y) - support(X) * support(Y) \tag{5}$$

- *Lift/Interest of a rule:* The lift value of a rule, $X \Rightarrow Y$, measures how many times more often X and Y occur together than expected if they were statistically independent. It ranges from $[0, +\infty]$ where a lift value of 1 indicates independence between X and Y, and higher values indicate a co-occurrence pattern.

$$lift(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X) * support(Y)} = \frac{confidence(X \Rightarrow Y)}{support(Y)} \tag{6}$$

Given a set o transactions D the problem of mining association rules is to generate all association rules that have a support and confidence greater than the user-specified minimum support and minimum confidence. Finding those rules is not trivial because of its combinatorial explosion.

The traditional association rule problem is an implication of the form $X \Rightarrow Y$ where X and Y appear in the same transaction (same time). The traditional problem has several extensions some of them presented in the book *"Top 10 algorithms in data mining"* [19]. One extension could be the problem of mining rules where X tends to appear along Y within a specific time. In this case we have to add time constraints to the antecedent and consequent. This problem is called *temporal association rule mining.*

For solving the association rules problem there exist several efficient algorithms like FP-growth or ECLAT. In this section, we will present the well-known Apriori algorithm for efficiently finding association rules, because it is the one that has been extended the most with temporal aspects. Afterward, we will introduce two solutions for the temporal association problem. The research in this field was very extensive. These two approaches were selected for being the ones that best adapted to our problem.

2.1 Apriori

Apriori [1] is used for finding frequent itemsets using candidate generation. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The algorithm is based on the idea that if an itemset is not frequent, any of its superset is never frequent.

By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. The first step of the algorithm simply counts item occurrences collecting those items that satisfy the minimum support requirement. A subsequent pass, say pass k, consists of three phases:

1. The itemset L_{k-1} found in the last step is used to generate the candidate itemsets C_k.

2. The transaction set is scanned and the support of candidates C_k is counted.

3. Itemsets that satisfy the minimum support requirement are added to the result

To generate the candidate itemsets mentioned in the 1rst phase, the algorithm follows this two steps process:

1. Join step: Generate R_k initial candidates of frequent itemsets of size k taking the union of the two frequent itemsets of size $k-1$, P_{k-1}, Q_{k-1}, that have the first $k-2$ elements in common.

2. Prune step: Check if all the itemsets of size $k-1$ in R_k are frequent and generate C_k by removing those that do not pass this requirement from R_k.

Several extensions on the Apriori algorithm are presented in the paper *Mining Sequential Patterns* [2]. These extensions are oriented on obtaining patterns with sequential data, although the time variable is not introduced. We founded more interesting for our context the two projects presented in the following sections.

2.2 Handling different type of temporal data

The first work I want to mention is introduced in two papers: *Data mining with temporal abstractions: learning rules from time series*[15] and *Temporal data mining for the analysis of administrative healthcare data*[7]. The authors present a new algorithm oriented to the mining of Temporal Association Rules. This algorithm handles events with a temporal duration and events represented by single time points.

They define each item as an event, and a sequence of events as a time ordered succession of episodes. They represent each temporal episode through a tuple with five fields:

- *Start time*
- *End time*
- *Subject:* The patient.
- *Variable group:* Hospital admissions (HA), ambulatory visits (AV), Drugs (Dr).
- *Variables:*
  - HA group: Diagnosis-related group, diagnoses and procedures.
  - AV group: ambulatory visits.
  - Dr group: drug prescriptions.

They define a Temporal Association Rule (TAR) as a relationship defined through a temporal operator which holds between an antecedent, consisting in one or more patterns, and a consequent, consisting in a single pattern. This temporal operators are defined through seven temporal operators; six of them are derived from Allen's algebra

[3] (*before, meets, overlaps, finished-by, equals, starts*). The last one is the more general *precedes* operator, which synthesizes all the six mentioned Allen's operators.

They also define three design parameters that will allow selecting only a subset of desired relationships. Given two episodes *e1* and *e2*, those parameters are the *left shift* (LS), defined as the maximum allowed distance between *e1.start* and *e2.start*, the *gap* (G), defined as the maximum allowed distance between *e1.end* and *e2.start*, and the *right shift* (RS), similarly defined as the maximum allowed distance between *e1.end* and *e2.end*.

Using the explained context, their method for TAR extraction is based on the Apriori strategy:

1. Iterative selection of a variable as consequent of the rule.
2. Extraction of the basic set of rules that fulfill the *precedes* temporal relationship and the *LS*, *RS* and *gap* parameters explained above. This resulting set contains all the rules with single cardinality in the antecedent.
3. Extraction of complex rules, defined as rules with antecedent of multiple cardinality K obtained through the intersection of the episodes of the antecedents of the rules of cardinality K-1

For each rule they compute the *support* and *confidence*. They define the *support* of a rule as the proportion of subjects for which the rule is verified over the total number of subjects involved in the study. They define the *confidence* of a rule as the ratio between the number of episodes of the antecedent involved in the rule and the total number of episodes of the antecedent during the whole observation period.

They apply the described framework to a large amount of data concerning all the main healthcare expenditures, including hospital admissions, pharmacological prescriptions, and ambulatory visits of the population of the Pavia (Italy) area. The results show that the framework is a useful method to observe frequent health care temporal patterns in a population, with the opportunity to monitor how the healthcare processes are working and to evaluate the compliance of the processes with respect to the recommended medical guidelines.

2.3 Temporal associations rules with multiple antecedents

The second interesting approach is presented in the paper *A Fast Algorithm for Mining Temporal Association Rules Based on a New Definition*[21] by Zhan, Li; Yu, Fusheng and Zhang, Huixin. The authors reform the definition of traditional association rules and then give a general form of temporal association rule. Based on the new definition, they propose a fast algorithm for mining temporal association rules.

Let $\mathcal{J} = \{I_1, I_2, ..., I_m\}$ be a set of items and $\mathcal{D} = \{(S_1, t_1), (S_2, t_2), ..., (S_n, t_n)\}$ be a temporal transaction dataset. They define a temporal association rule as an implication of the form $X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{p-1}}{\Rightarrow} X_p$ $(p \geq 2)$, such that $X_k \subset \mathcal{J}, k = 1, 2, ...p$ and $(t_1, t_2, ...t_{p-1})$, is called *time constraint*. Every t_i is a given positive integer. $X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{p-2}}{\Rightarrow} X_{p-1}$ is called *antecedent* while X_p is called *consequent*.

In their proposal, a temporal association rule $X \overset{t}{\Rightarrow} Y$ doesn't follow the traditional association rules condition of $X \cap Y = \emptyset$. They propose the following formulation:

Let

$$g_{S_i} : \mathcal{P}(\mathcal{J})^p \rightarrow \{0, 1\}, \qquad g_{S_i}(X) = \begin{cases} 1 & \text{if } i \text{ contains } X \\ 0 & \text{else} \end{cases} \tag{7}$$

Then

$$g^t_{S_i}(X_1, ..., X_p) = \min(g_{S_{i_1}}(X_1), ...g_{S_{i_p}}(X_p)) \tag{8}$$

Let

$$h^t : \mathcal{P}(\mathcal{J})^p \rightarrow \mathbb{N}, \qquad h^t(X_1, X_2, ..., X_p) = \sum_{S_i \in D} g^t_{S_i}(X_1, X_2, ..., X_p) \tag{9}$$

Then the support and confidence of $X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{p-1}}{\Rightarrow} X_p$ are defined as follows.

$$\textit{support } (X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{p-1}}{\Rightarrow} X_p) = \frac{h^t(X_1, X_2, ..., X_p)}{h^t(\emptyset, ..., \emptyset)} \tag{10}$$

$$\textit{confidence } (X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{p-1}}{\Rightarrow} X_p) = \frac{\textit{support}(X_1 \overset{t_1}{\Rightarrow} ... \overset{t_{p-2}}{\Rightarrow} X_{p-1} \overset{t_{p-1}}{\Rightarrow} X_p)}{\textit{support}(X_1 \overset{t_1}{\Rightarrow} ... \overset{t_{p-2}}{\Rightarrow} X_{p-1} \overset{t_{p-1}}{\Rightarrow} \emptyset)} \tag{11}$$

There is another temporal association rule in the form of $X_1 \overset{T_1}{\Rightarrow} X_2 \overset{T_2}{\Rightarrow} ... \overset{T_{p-1}}{\Rightarrow} X_p$, where $T_i = [t_{i_1}, t_{i_2}]$ with $t_{i_1} < t_{i_2}$. This implies that the itemset X_{i+1} is purchased between the t_{i_1}-th unit time and t_{i_2}-th unit time after X_i purchased. Then, the support and confidence of the temporal association rule are defined as follows.

$$\textit{support } (X_1 \overset{T_1}{\Rightarrow} ... \overset{T_{p-1}}{\Rightarrow} X_p) = \frac{\sum_{t_1 \in T_1} \cdots \sum_{t_{p-1}} h^{(t_1, ..., t_{p-1})}(X_1, ..., X_p)}{\sum_{t_1 \in T_1} \cdots \sum_{t_{p-1}} h^{(t_1, ..., t_{p-1})}(\emptyset, ..., \emptyset)} \tag{12}$$

$$\textit{confidence } (X_1 \overset{T_1}{\Rightarrow} ... \overset{T_{p-1}}{\Rightarrow} X_p) = \frac{\textit{support}(X_1 \overset{T_1}{\Rightarrow} ... \overset{T_{p-2}}{\Rightarrow} X_{p-1} \overset{T_{p-1}}{\Rightarrow} X_p)}{\textit{support}(X_1 \overset{T_1}{\Rightarrow} ... \overset{T_{p-2}}{\Rightarrow} X_{p-1} \overset{T_{p-1}}{\Rightarrow} \emptyset)} \tag{13}$$

Using this framework they propose the following algorithm:

1. Find out all those item sets in $\{S_1, S_2, ..., S_n\}$ whose support is not less than $\dfrac{n - \sum_{i=0}^{p-1} t_i}{n} \times minsup$ by Apriori Algorithm. Denotate those by F_1. Let $k = 2$.

2. Take F_{k_1} as antecedent and F_1 as consequent.Find out all those temporal association rules of form $X_1 \overset{t_1}{\Rightarrow} X_2 \overset{t_2}{\Rightarrow} ... \overset{t_{k-1}}{\Rightarrow} X_k$ whose support is not less than $\dfrac{n - \sum_{i=0}^{p-1} t_i}{n - \sum_{i=0}^{k-1} t_i} \times minsup$. Denote those temporal association rules by F_k. Find out those rules which have *minsup* and *minconf* in F_k and denote those rules by R_k.

3. If $F_k = \emptyset$ or $k \geq p$, go to Step 4; otherwise let $k++$ and repeat Step 2.

4. $\bigcup_{i=2}^{k} R_k$ has all temporal association rules.

They test the algorithm with synthetic and a real datasets. The experiments exhibit the good performance of the new proposed algorithm. The proposed algorithm can find out temporal association rules effectively.

3 Methodology

In this section we are going to specify our methodology that is going to be applied in this project. Firstly, all the steps included in our workflow will be described. Also, all the technologies used in the project will be listed.

3.1 Strategy

The life cycle of this project is based on the well-known CRISP-DM standard process model. The methodology differences six main steps in a data mining process: Problem understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Implementation.

- **Problem Understanding:** Consists on determining the problem, analyze the state of the art, and create the data mining objectives. We are aware that these objectives are mutable depending on the feedback received by the doctors. The objectives will also be adapted to the time-boundary of a master thesis.
- **Data Understanding:** Traditionally consists of a data collection phase and continues with the activities that allow you to familiarize yourself with the data, identifying quality problems, and discover preliminary knowledge about the data. In our case, the data collection phase does not exist as it was done in another project. This makes more important other mentioned phases, as we need to understand the work done by another student.
- **Data Preparation:** In this phase, we will prepare our data for the modeling section. In our case, the steps applied will mainly be focused on performing some modifications to string data, and generalizing the medical codes while correcting some errors detected in the data collection phase.
- **Modeling:** Modeling techniques that are relevant to the problem are selected and applied. In our case, we will apply the formulation of our temporal apriori-based algorithm that we defined at the beginning when analyzing the state of the art.
- **Evaluation:** This phase could be divided into two steps: A self-evaluation reviewing the steps taken to create the results and comparing the results with other executions, and an external evaluation where we will show the obtained results to some medical professionals.
- **Implementation:** This final phase is to present the tool/knowledge to the client so that they can use it. This phase will be out of our project for time limitations.

The sequence of the phases is not strict, moving back and forth between different steps is always required. During the execution of this project, several iterations of this process were applied, performing different preprocessing steps, reformulating our temporal association rules framework, and searching for optimizations. As a general concept, we tried to build a first simple version of a code and perform iterations over it.

3.2 Technologies

The following technologies will be used:

3.2.1 MyDisk

For privacy purposes, any document or code file related to the project can't be uploaded to GitHub, Google Drive, or another type of version-control platform. We worked with the MyDisk [12] cloud service platform, which is hosted at the UPC, and provides an easy-to-use, safe and reliable environment for research groups and collaborators.

3.2.2 Neo4J

Neo4j [13] is a free graph-oriented database software. Its developers describe it as an ACID (atomicity, consistency, isolation, durability) compliant transactional database with native graph storage and processing. Neo4J is very useful if for a given problem there is a special interest in the relations of the instances. The queries can be made using the Cypher [8] language. Cypher is a declarative, SQL-inspired language for describing visual patterns in graphs. We will use Neo4J because the medical records were already loaded in this database. We will use Neo4J version 4.2.1.

3.2.3 Python

Python [14] is an interpreted, object-oriented, high-level programming language. Python will be the main programming language of this project. The use of Python was a requirement of the project directors because the rest of the software of the TADIA-MED project is developed in this programming language. We will use Python 3.8.1.

4 Development

In this chapter, the available data for our project is summarily described, followed by the data extraction, data preparation, data preprocessing steps that we applied. Finally, our formulation will be described.

4.1 Available data

As we mentioned before, IDIAP Jordi Gol is the property of the data that will be used in the project. The data contains information on patient visits at Primary Care Centers (CAP) throughout Catalonia. We have anonymized information on multimorbid patients that were visited between 2010 to 2016. The "multimorbid" term means that the patients must have in their medical record at least one of these symptoms. The symptom must have happened when the patient was older than 50 years old:

- Accidente Isquémico Transitorio
- Otros tipos de Accidentes Vasculares Cerebrales (AVC)
- AVCs hemorrágicos
- AVCs isquémicos
- Secuelas provocadas por AVC
- Cáncer de pulmón
- Cáncer de colon
- Infarto agudo de miocardio
- Defunción

Our corpus was built with the information of these multimorbid of patients. All annotations that the doctor registers in the health system during the visits follow what is called the MEAP structure (**M**otiu de la consulta, **E**xploració, **A**valuació i **P**la d'actuació, in Catalan). The information contained in these MEAPs was manually annotated with the 4 types of nodes and relationships mentioned later, producing different *xml* documents. These documents were imported into a Neo4J database. Since the corpus was manually annotated, it can contain inconsistencies (as mentioned later, in Section 4.4.2).

The Neo4J server is shared with different students. The server contains data structured as shown in Figure 1. The figure shows information coming from two different sources, that generate a graph with two connected components. MIMIC's database containing information relating to patients who stayed within the intensive care units at Beth Israel Deaconess Medical Center [5]. This information is outside the scope of our project, we will focus on the IDIAP connected component.

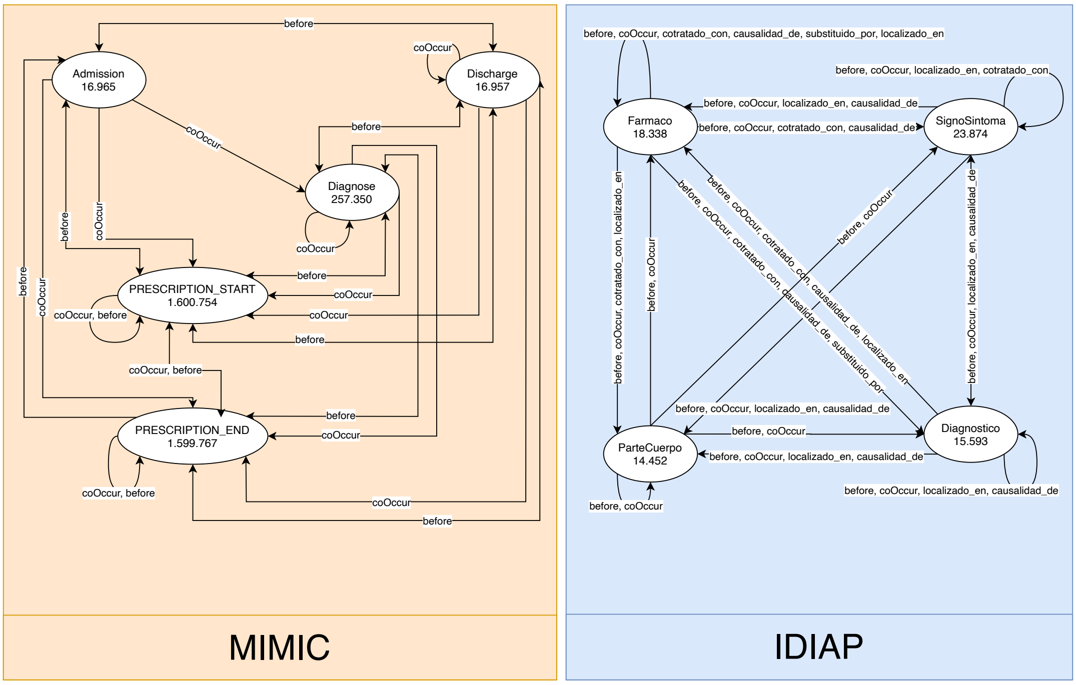

before
coOccur
Admission
16.965
Discharge
16.957
coOccur
before
Diagnose
257.350
coOccur
before
coOccur
before
PRESCRIPTION_START
1.600.754
coOccur, before
before
coOccur
coOccur
before
coOccur, before
before
coOccur
PRESCRIPTION_END
1.599.767
coOccur, before
coOccur
before
MIMIC
before, coOccur, cotratado_con, causalidad_de, substituido_por, localizado_en
before, coOccur, localizado_en, cotratado_con
before, coOccur, localizado_en, causalidad_de
Farmaco
18.338
before, coOccur, cotratado_con, causalidad_de
SignoSintoma
23.874
before, coOccur, cotratado_con, localizado_en
before, coOccur
before, coOccur, cotratado_con, causalidad_de, substituido_por
before, coOccur, cotratado_con, causalidad_de, localizado_en
before, coOccur
before, coOccur, localizado_en, causalidad_de
before, coOccur, localizado_en, causalidad_de
before, coOccur
ParteCuerpo
14.452
before, coOccur, localizado_en, causalidad_de
Diagnostico
15.593
before, coOccur, localizado_en, causalidad_de
before, coOccur
IDIAP

Figure 1: Data model

The IDIAP component contains four types of nodes: *Farmaco* (Drug), *SignoSintoma* (Sign or symptom), *ParteCuerpo* (Part of the body), and *Diagnostico* (Diagnosis). We also can observe that there are six types of edges: *before*, *coOccur*, *causalidad_de* (causality), *cotratado_con* (treated at the same time with), *localizado_en* (located in), *substituido_por* (substituted by). There are no properties stored in the edges. We are first going to analyze the properties of each type of node, and afterward, how the nodes can be related.

4.1.1 Node properties

4.1.1.1 Farmaco

Attribute	Description	Example/Possible values
raw_text	Doctor's note	ex: Lexemma
date	Date of the note	ex: 20150217
code	Type of code and code	ex: atc7:D07AC14
tiempo	Moment in which the diagnosis occurs	P: Past A: Actuality
patient_id	Patient identifier	ex: 345

id_node	Node identifier	ex: 79929
atc7	Most accurate ATC code that can be related to the drug	ex: D07AC14
cert	Certainty	N: Negation P: Possibility A: Affirmation
text	Description based on the code	ex: metilprednisolona, aceponato de
id	Node identifier within each patient. Follows temporal order.	ex: 756
dataset	Dataset where the data comes from	ex: idiap
sit	Drug status	A: Start the prescription B: End the prescription

Table 1: Available data for Farmaco

The Anatomical Therapeutic Chemical (ATC) Classification System [4] is a drug classification system controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC). The ATC codes have different levels of hierarchy:

- The first level of the code indicates the anatomical main group. It consists of one letter.
- The second level of the code indicates the therapeutic subgroup. It consists of two digits.
- The third level of the code indicates the therapeutic/pharmacological subgroup. It consists of one letter.
- The fourth level of the code indicates the chemical/therapeutic/pharmacological subgroup. It consists of one letter.
- The fifth level of the code indicates the chemical substance. It consists of two digits.

In the Figure 2 we can observe a summary of the generalization tree of the *atc7* code and our definition of the different level numbers. This definition will be used in future sections. We can observe that despite the maximum tree height is five, not all nodes have the same number of children and that there's not the same distance from all the leaves to the root node. But we can observe that all the codes belonging to the same level follow the same structure, as mentioned before. These three things will be important when generalizing the *atc7* code.

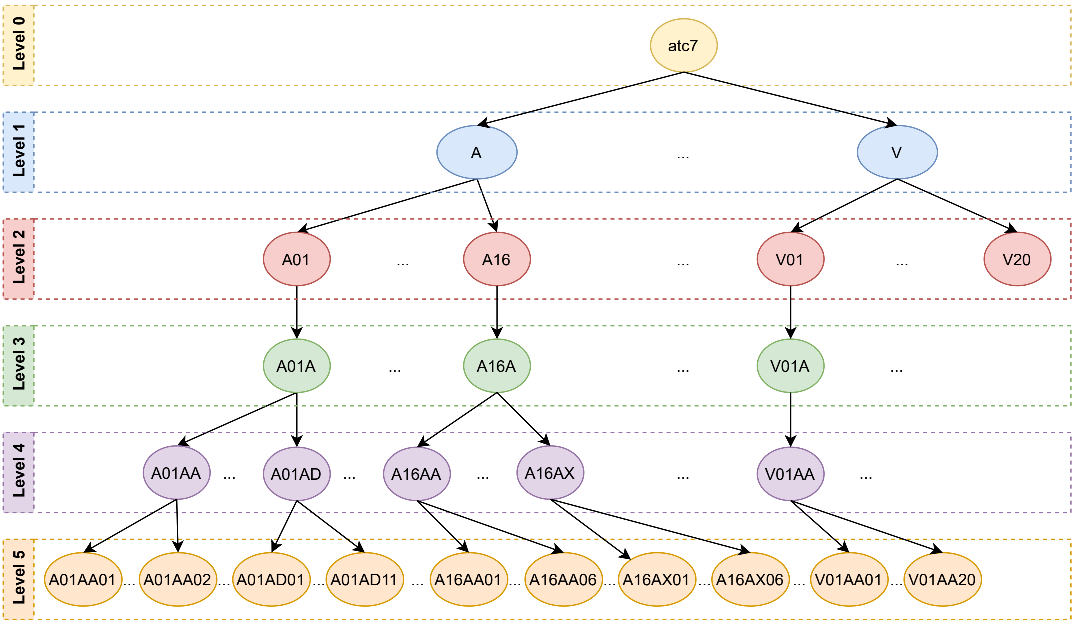


Figure 2: ATC7 generalization tree summary

4.1.1.2 SignoSintoma

Attribute	Description	Example/Possible Values
raw_text	Doctor's note	ex: odinofagia
date	Date of the note	ex: 20150217
code	Type of code and code	ex: ciap2:D21
tiempo	Moment in which the sign or symtom is detected	P: Past A: Actuality
patient_id	Patient identifier	ex: 345
id_node	Node identifier	ex: 80635
cert	Certainty	N: Negation P: Possibility A: Affirmation
text	Description based on the code	ex: PROBLEMAS DE LA DEGLUCIÓN
id	Node identifier within each patient. Follows temporal order.	ex: 761
dataset	Dataset where the data comes from	ex: idiap

Table 2: Available data for SignoSintoma

The CIAP-2 codes (Clasificación Internacional de Atención Primaria) is a three digit code were:

- The first character is a letter that represents organic system. There are 17 possible values.
- The second and third characters are digits that represent the "components", which are related specifically or nonspecifically with the reason for consultation, illness or health problem.

In the Figure 3 we can observe a summary of the generalization tree of the *ciap2* code and our definition of the different level numbers. The image is a summary but it doesn't mean that all the nodes have the same number of children, but contrary to what happens in the *atc7* code, in *ciap2* there's the same distance from between all the leaves and the root node. This distance is two.

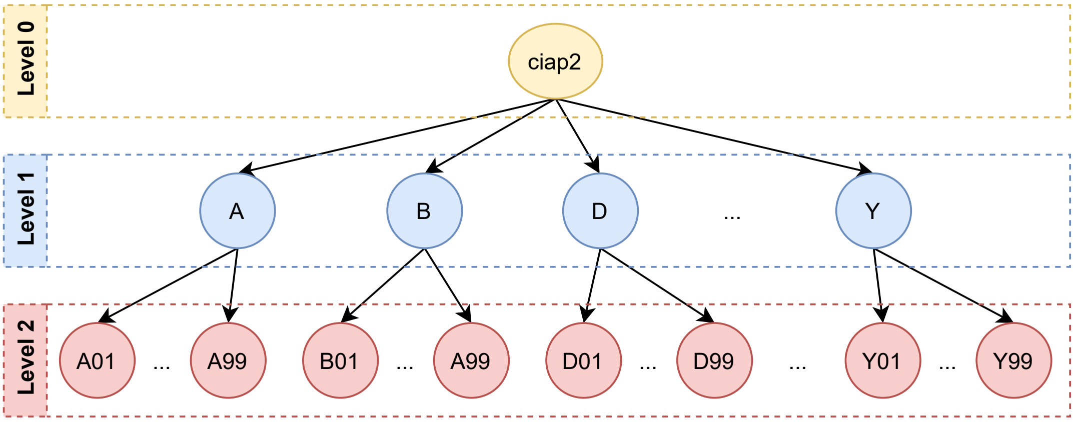


Figure 3: CIAP2 generalization tree summary

4.1.1.3 ParteCuerpo

Attribute	Description	Example/Possible Values
raw_text	Doctor's note.	ex: ABDOMINAL
date	Date of the note.	ex: 20150102
code	Type of code and code.	ex: snomed:302553009
patient_id	Patient identifier.	ex: 345
id_node	Node identifier.	ex: 80054
text	Description based on the code.	ex: ABDOMINAL
id	Node identifier within each patient. Follows temporal order.	ex: 755
dataset	Dataset where the data comes from	ex: idiap

Table 3: Available data for ParteCuerpo

SNOMED [16] (Systematized Nomenclature of Medicine) is a clinical reference terminology that enables healthcare professionals around the world to represent clinical information accurately and unambiguously, in a multilingual format. In this project will not distinguish any kind of hierarchy in the *snomed* codes.

4.1.1.4 Diagnostico

Attribute	Description	Example/Possible values
raw_text	Doctor's note	ex: hipetrofia de lobulo hepatico izquierdo
Date	Date of the note	ex: 20141231
cim10	More accurate cim10 code that can be related to diagnosis.	ex: K76.0
code	Type of code and code	ex: cim10:K76.0
tiempo	Moment in which the diagnosis occurs	P: Past A: Actuality
patient_id	Patient identifier	ex: 345
id_node	Node identifier	ex: 81835
cert	Certainty	N: Negation P: Possibility A: Affirmation
text	Description based on the code	ex: hígado graso (degeneración grasa) de hígado, no clasificado bajo otro concepto
id	Node identifier within each patient. Follows temporal order.	ex: 754
dataset	Dataset where the data comes from	ex: idiap

Table 4: Available data for Diagnostico

The CIM-10 code is a medical classification list created by the World Health Organization (WHO). It contains codes for diseases, signs, and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. We will only use the *cim10* classification for diagnoses. The code has a non-trivial hierarchy, which can be found on the official page of the CIM codes in Spanish [6]. In Figure 4 we try to show the complexity of the code and define the generalization levels.

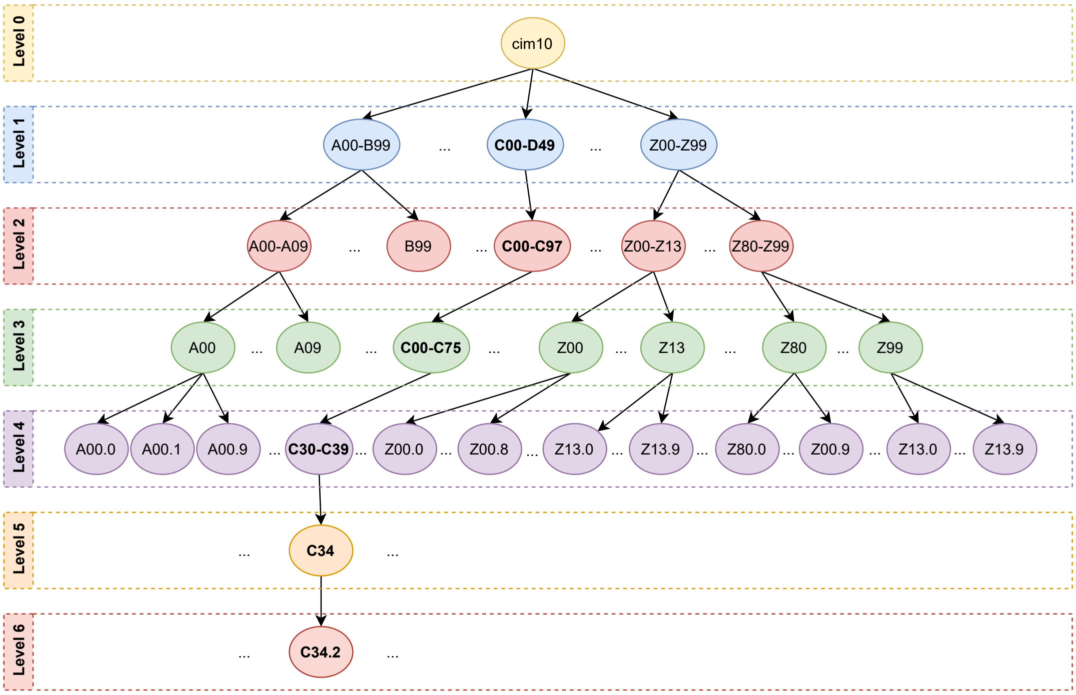


Figure 4: CIM10 generalization tree summary

When we say that the code has a non-trivial hierarchy is because not all the nodes have the same number of children, we don't have the same distance from all the leaves to the root node, and because the labels inside each level do not follow the same structure. In Figure 4, with bold labels, we wanted to show the largest path (among other paths). We can observe the tree height is six.

4.1.2 Interaction between nodes

As we mentioned before there are six types of edges: *before*, *coOccur*, *causalidad_de* (causality), *cotratado_con* (treated at the same time with), *localizado_en* (located in), *substituido_por* (substituted by).

The *coOccur* edge connects all pairs of nodes that belong to the same visit. We define a visit as all the instances that have the same *patient_id* and *date* properties. With this definition, we assume that the patients can only have one visit per day. The *before* edge connects all pairs of nodes that belong to two consecutive visits. The *coOccur* and *before* edge connections can be observed in Figure 5, where we have 4 visits for patient 31: day 20161109 (in the red circle), day 20161110 (in the orange circle), day 20161111 (in the yellow circle) and day 20161112 (in the green circle).

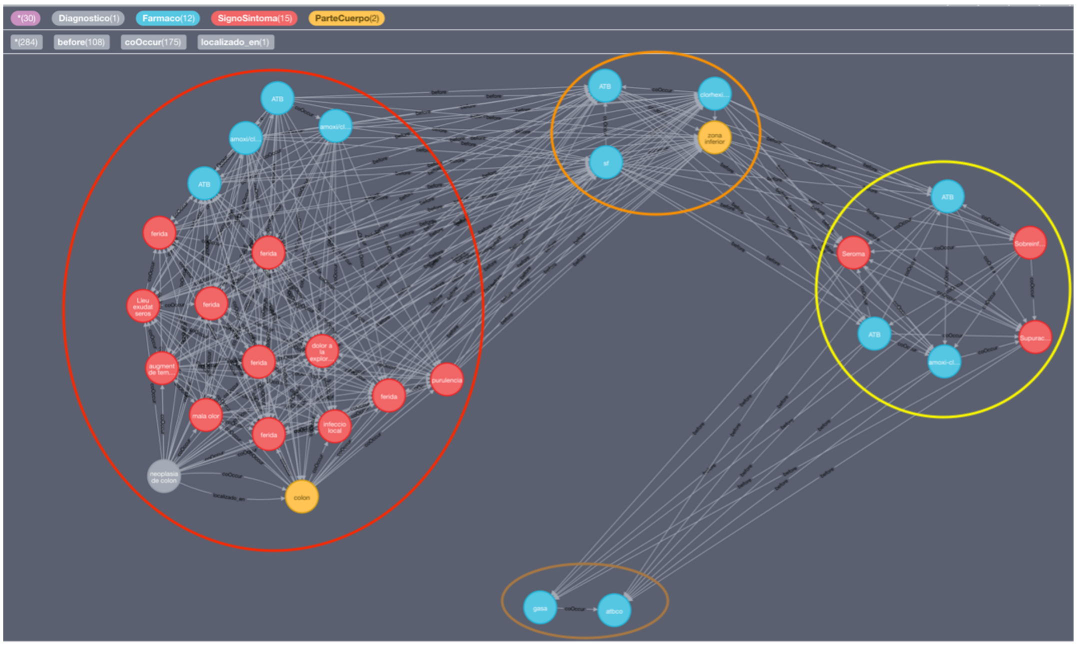
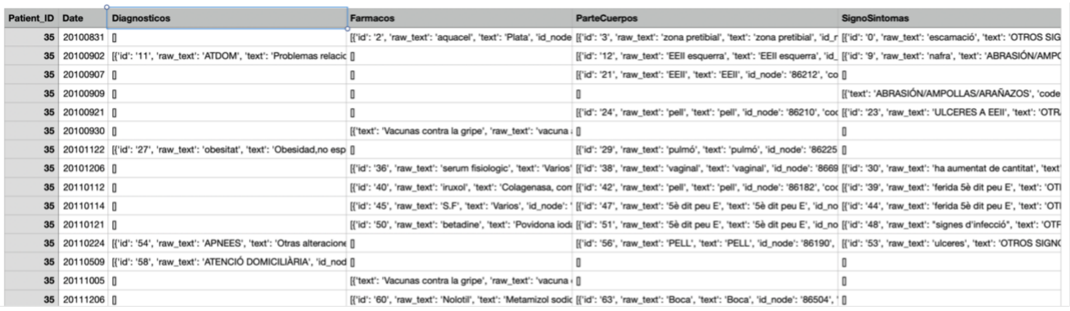


Figure 5: Instances from patient 32, days 20161109, 20161110, 20161111 and 20161112

Regarding the other relations, they give special information to a given *before* or *coOccur* relation. But if a *causalidad_de*, *cotratado_con*, *localizado_en* or *substituido_por* relation between two nodes exists, the instances will also be connected by one of the two temporal edges (*before* or *coOccur*). The only exception of this rule is if the relation is a recursive relation, then the temporal connection does not exist.

4.2 Data extraction

Once understood how the data was structured in the Neo4J server, we wanted to extract the data into a *.csv* file to proceed with the exploratory analysis. We build a *python* script that executed some *cypher* queries that obtained all the nodes of each visit in the Neo4J server. We considered the *patient_id* and *date* the identifiers of each row of the *.csv* file. Besides the identifiers, each row of the file also contains a list of all the *Diagnostico*, *Farmaco*, *ParteCuerpo* and *SignoSintoma* nodes involved in each visit. For each node we stored all the information explained in Section 4.1.1 of this document. Figure 6 shows a capture of the mentioned *.csv* file.



Patient_ID	Date	Diagnosticos	Farmacos	ParteCuerpos	SignoSintomas
35	20100831	[]	[{'id': '2', 'raw_text': 'aquacel', 'text': 'Plata', 'id_node	[{'id': '3', 'raw_text': 'zona pretibial', 'text': 'zona pretibial', 'id_r	[{'id': '0', 'raw_text': 'escamació', 'text': 'OTROS SIG
35	20100902	[{'id': '11', 'raw_text': 'ATDOM', 'text': 'Problemas relacic	[]	[{'id': '12', 'raw_text': 'EEII esquerra', 'text': 'EEII esquerra', 'id_	[{'id': '9', 'raw_text': 'nafra', 'text': 'ABRASIÓN/AMPC
35	20100907	[]	[]	[{'id': '21', 'raw_text': 'EEII', 'text': 'EEII', 'id_node': '86212', 'co	[]
35	20100909	[]	[]	[]	[{'text': 'ABRASIÓN/AMPOLLAS/ARAÑAZOS', 'code
35	20100921	[]	[]	[{'id': '24', 'raw_text': 'pell', 'text': 'pell', 'id_node': '86210', 'coc	[{'id': '23', 'raw_text': 'ULCERES A EEII', 'text': 'OTR
35	20100930	[]	[{'text': 'Vacunas contra la gripe', 'raw_text': 'vacuna	[]	[]
35	20101122	[{'id': '27', 'raw_text': 'obesitat', 'text': 'Obesidad,no esp	[]	[{'id': '29', 'raw_text': 'pulmó', 'text': 'pulmó', 'id_node': '86225	[]
35	20101206	[]	[{'id': '36', 'raw_text': 'serum fisiologic', 'text': 'Varios'	[{'id': '38', 'raw_text': 'vaginal', 'text': 'vaginal', 'id_node': '8669	[{'id': '30', 'raw_text': 'ha aumentat de cantitat', 'text'
35	20110112	[]	[{'id': '40', 'raw_text': 'iruxol', 'text': 'Colagenasa, con	[{'id': '42', 'raw_text': 'pell', 'text': 'pell', 'id_node': '86182', 'coc	[{'id': '39', 'raw_text': 'ferida 5è dit peu E', 'text': 'OTI
35	20110114	[]	[{'id': '45', 'raw_text': 'S.F', 'text': 'Varios', 'id_node': '	[{'id': '47', 'raw_text': '5è dit peu E', 'text': '5è dit peu E', 'id_no	[{'id': '44', 'raw_text': 'ferida 5è dit peu E', 'text': 'OTI
35	20110121	[]	[{'id': '50', 'raw_text': 'betadine', 'text': 'Povidona iod	[{'id': '51', 'raw_text': '5è dit peu E', 'text': '5è dit peu E', 'id_no	[{'id': '48', 'raw_text': "signes d'infecció", 'text': 'OTF
35	20110224	[{'id': '54', 'raw_text': 'APNEES', 'text': 'Otras alteracione	[]	[{'id': '56', 'raw_text': 'PELL', 'text': 'PELL', 'id_node': '86190',	[{'id': '53', 'raw_text': 'ulceres', 'text': 'OTROS SIGNC
35	20110509	[{'id': '58', 'raw_text': 'ATENCIÓ DOMICILIÀRIA', 'id_nod	[]	[]	[]
35	20111005	[]	[{'text': 'Vacunas contra la gripe', 'raw_text': 'vacuna	[]	[]
35	20111206	[]	[{'id': '60', 'raw_text': 'Nolotil', 'text': 'Metamizol sodic	[{'id': '63', 'raw_text': 'Boca', 'text': 'Boca', 'id_node': '86504', '	[]

Figure 6: CSV file obtained after extracting the data

4.3 First exploratory data analysis

The IDIAP dataset contains information about 320 patients, with a total number of 72,257 healthcare episodes, distributed as shown in Table 5.

Node	N variables code	N variables raw_text	N episodes
Farmaco	796	3742	18338
SignoSintoma	249	10455	23874
ParteCuerpo	601	5184	14452
Diagnostico	1193	6429	15593
Total	2839	25810	72257

Table 5: Episodes and Variables distribution before preprocessing

We wanted to analyze the type of values inside the *raw_text* and *code* variables. Table 6 - 9 show the most frequent values for each variable in each type of node. We want to highlight two aspects: The first one is that, as we expected, there's some *raw_text* values that from our non expert point of view, represent the same concept. An example can be `disnea` and `dispnea` or `fiebre` and `febre`. In the preprocessing step we will apply some string distance in order to reduce de variability in the *raw_text* attribute. The second thing we can observe is that not all the codes are expressed in its lower level of generalization. For example, the code `atc7:A10`, which refers to `insulina`, could be more concrete, for example with the code `atc7:A10AB04`, that represents `insulina lispro`. We will have to take this into account when generalizing the *code* variable.

	Name	Count	%
0	paracetamol	399	2.18
1	sintrom	324	1.77
2	insulina	288	1.57
3	TAO	275	1.50
4	enalapril	186	1.01
5	diclofenaco	181	0.99
6	Paracetamol	167	0.91
7	AAS	146	0.80
8	Sintrom	133	0.73
9	omeprazol	129	0.70
10	furosemida	127	0.69
11	VAG	126	0.69
12	ATB	113	0.62
13	SF	112	0.61
14	ibuprofeno	110	0.60

	Name	Count	%
0	N02BE01	766	4.18
1	B01AA07	583	3.18
2	M01AB05	494	2.69
3	A10	357	1.95
4	C09AA02	351	1.91
5	N02BB02	348	1.90
6	C01EB16	316	1.72
7	C03CA01	316	1.72
8	J01	311	1.70
9	B01	311	1.70
10	B03BA01	276	1.51
11	D08AL30	275	1.50
12	A02BC01	274	1.49
13	J07BB	257	1.40
14	J01CR02	225	1.23

Table 6: 10 most frequent values for Farmaco nodes

	Name	Count	%
0	dolor	736	3.08
1	tos	458	1.92
2	MVC	380	1.59
3	fiebre	320	1.34
4	BEG	279	1.17
5	febre	248	1.04
6	disnea	244	1.02
7	mvc	240	1.01
8	febril	159	0.67
9	soplos	153	0.64
10	dispnea	141	0.59
11	NH	126	0.53
12	NC	113	0.47
13	diarrea	107	0.45
14	roncus	99	0.41

	Name	Count	%
0	A29	2125	8.90
1	A01	1725	7.23
2	K29	1345	5.63
3	R29	1145	4.80
4	N29	975	4.08
5	A03	967	4.05
6	R02	864	3.62
7	R05	844	3.54
8	D29	689	2.89
9	R04	686	2.87
10	S29	546	2.29
11	L29	493	2.07
12	S19	416	1.74
13	K05	358	1.50
14	A04	312	1.31

Table 7: 10 most frequent values for SignoSintoma nodes

	Name	Count	%
0	EEII	323	2.23
1	abdominal	309	2.14
2	lumbar	194	1.34
3	Abdomen	141	0.98
4	pulmonar	136	0.94
5	colon	117	0.81
6	ABD	105	0.73
7	abdomen	104	0.72
8	cames	92	0.64
9	cervical	77	0.53
10	faringe	70	0.48
11	peus	70	0.48
12	OI	70	0.48
13	pell	69	0.48
14	OD	67	0.46

	Name	Count	%
0	243996003	1230	8.51
1	302553009	966	6.68
2	302551006	599	4.14
3	1910005	443	3.07
4	181216001	440	3.04
5	302545001	365	2.53
6	182343007	316	2.19
7	182083008	276	1.91
8	302541005	251	1.74
9	450807008	241	1.67
10	244486005	239	1.65
11	181817002	229	1.58
12	181469002	219	1.52
13	302536002	216	1.49
14	302538001	213	1.47

Table 8: 10 most frequent values for ParteCuerpo nodes

	Name	Count	%
0	HTA	526	3.37
1	DM	243	1.56
2	PADES	211	1.35
3	IAM	154	0.99
4	AVC	150	0.96
5	ATDOM	138	0.89
6	defunció	130	0.83
7	ferida	128	0.82
8	ITU	127	0.81
9	MPOC	115	0.74
10	DM2	109	0.70
11	CVA	103	0.66
12	fumador	98	0.63
13	anemia	88	0.56
14	DLP	86	0.55

	Name	Count	%
0	I10	739	4.74
1	E14	440	2.82
2	J00	351	2.25
3	I64	310	1.99
4	Z76.8	302	1.94
5	F17.1	263	1.69
6	Z74	258	1.65
7	E11	252	1.62
8	T14.1	249	1.60
9	J44.9	237	1.52
10	I46.9	226	1.45
11	H26.9	225	1.44
12	I48	217	1.39
13	I21.9	205	1.31
14	N39.0	201	1.29

Table 9: 10 most frequent values for Diagnostico nodes

Finally, regarding the visits of a patient concept, we computed some base statistics that will help to understand our data to parametrize our algorithm in future sections. We have a total of 14,330 visits. The patients have a mean of 44.78 visits. The patient that has more visits has 205, and the minimum visits per patient are 1. The visits have a mean of 5.04 nodes involved. The mean time between two consecutive visits of the same patient is 42.83 days.

4.4 Data preparation

4.4.1 Preprocessing strings

After performing the first exploratory data analysis, we dealt with the main two problems we discovered in the Exploratory Data Analysis. Regarding the *raw_text* variable, we applied the following preprocessing steps:

- Case folding: We converted all the letters of the strings to lowercase letters.
- Stopwords removal: We removed the prepositions and articles. We finally kept the adverbs because some of them are relevant.
- String distance: We applied the Levenshtein distance to those *raw_text* that have more than three characters. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other. We grouped all the values that had distance equal or smaller than two.

4.4.2 Generalization of codes

We also applied different levels of generalization to each type of *code*, and stored the results in different *.csv* files. The *Farmaco* and *SignoSintoma* nodes have an easy generalization: We cut the string for the part we are interested in, taking into account the code hierarchy described in Section 4.1.1. Regarding the *Diagnostico* nodes, as the hierarchy is not-trivial, we used the `cie` python library[11] that scraped the information from the official page of the codes in Spanish[6]. We do not distinguish different levels of generalization in the *ParteCuerpo* nodes, so we will not apply any change to them.

After exploring the data, we have decided to work with two different levels of generalization. The first one is with all the codes at the level that they appear in the dataset, without any generalization. This is the same as applying the following levels of generalization:

- *Diagnostico:* Level 6
- *Farmaco:* Level 5
- *ParteCuerpo:* Lets say X, we are not generalizing this node.
- *SignoSintoma:* Level 2

We will name this generalization *65X2*. We also will work with a medium-level of generalization:

- *Diagnostico:* Level 2
- *Farmaco:* Level 3
- *ParteCuerpo:* Lets say X, we are not generalizing this node.
- *SignoSintoma:* Level 2

We will name this generalization *23X2*. When we mention a code level, we are specifying the maximum code level allowed. However, if a given node does not have the specified level, we will use its maximum generalized level. For example, in Figure 2 we can observe that the label *V20* can't be generalized to the third level of the *atc7* codes, so we will use its second level generalization.

Another problem found is that some codes are wrong codified. For those codes, we will codify them in a higher level of abstraction, if possible. Table 10 and 11 shows the errors found and the corrected code applied. In Table 11 there are some values that, with the given information we could not deduce any *ciap2* code. We will leave them as they are and if the code is used, it will be discarded by the minimum support.

	Wrong Code	Correction	N instances
0	C20.9	C20	1
1	C78.9	C78	5
2	C80.9	C80	5
3	F42.3	F42	3
4	G20.9	G20	1
5	I50.90	I50	1
6	I50.91	I50	6
7	M12.7	M12	3
8	M13.2	M13	8
9	M45.9	M45	3
10	R47.9	R47	1
11	Z13.60	Z13	22

Table 10: Errors found when generalizing CIM10 code

	Wrong Code	Correction	N instances
0	text:acufenos	ciap2:H03	5
1	text:acúfens	ciap2:H03	3
2	text:acufens	ciap2:H03	3
3	text:estenosis	None	2
4	cim10:S29	ciap2:L04	2
5	text:Vesq moderadament hipertrofic	None	2
6	text:ACUFENOS	ciap2:H03	2
7	text:polipos colon	ciap2:D29	1
8	text:acúfenos	ciap2:H03	1
9	cim10:N29	ciap2:U14	1
10	text:ull vermell	ciap2:F02	1
11	text:Ull vermell	ciap2:F02	1
12	text:Fase terminal	ciap2:A96	1
13	text:infeccions	None	1
14	text:Pàncrees amagat per els quistes	ciap2:D99	1
15	text:ACUFENS CRÒNICS NOCTURNS	ciap2:H03	1
16	text:FRAGILITAT	ciap2:A04	1
17	text:Fragilitat	ciap2:A04	1
18	cim10:S19	ciap2:L01	1
19	cim10:R29	ciap2:N73	1
20	text:ESGOTADA	ciap2:A04	1
21	text:Acufenos	ciap2:H03	1
22	text:ACUFENS	ciap2:H03	1
23	text:derrame pleural	ciap2:R82	1
24	text:acufeno en OD	ciap2:H03	1
25	text:masa a recte	ciap2:D75	1
26	cim10:K28	ciap2:D86	1
27	text:frecs	ciap2:K05	1
28	text:estenosis leve en ACI derecha	ciap2:K99	1
29	text:dismorfia septal	ciap2:D20	1
30	text:signos de HVizq	ciap2:K29	1
31	text:disfunció ventricular severa	ciap2:K05	1
32	cim10:M21.7	ciap2:L29	1
33	cim10:J00	ciap2:A29	1
34	cim10:M54.2	ciap2:L01	1

Table 11: Errors found when generalizing CIAP2 code

We decided to use the second generalization level to *SignoSintoma* nodes because the first level is too wide for interpreting the associations. Regarding the third level in *Farmaco* nodes is based on the results found in the paper *Data mining with temporal abstractions: learning rules from time series*[15]. Finally, in the *Diagnostico* node we obtained results with several code levels. We finally decided to keep the second level because it's a good tradeoff between interpretability and generalization.

4.5 Second exploratory Data Analysis

After all the preprocessing steps applied, we wanted to analyze the type of values inside the *raw_text* and *code* variables. Table 12 shows the number of values for the *code* and *raw_text* variables after preprocessing each generalization.

	Generalization 65X2		Generalization 23X2		
Node	N code	N raw_text	N code	N raw_text	N episodes
Farmaco	796	3742	202	1595	18338
SignoSintoma	249	10455	249	7100	23874
ParteCuerpo	601	5184	601	2960	14452
Diagnostico	1193	6429	588	3947	15593
Total	2839	25810	1640	15602	72257

Table 12: Episodes and Variables distribution after preprocessing

For each node we also wanted to observe most frequent *code* labels per each generalization and the modified *raw_text* variables. Tables 13 to 16 show the count and percentage per each variable. Table 16 is composed of 1 subtable because the codes for *ParteCuerpo* are modified neither in generalization *65X2* nor *23X2*.

				Level 5			Level 3		
	Name	Count	%	Name	Count	%	Name	Count	%
0	paracemtaol	666	3.63	N02BE01	766	4.18	B01A	1353	7.38
1	sintrom	575	3.14	B01AA07	583	3.18	N02B	1327	7.24
2	inssulina	353	1.92	M01AB05	494	2.69	M01A	794	4.33
3	enalnapril	339	1.85	A10	357	1.95	N02A	523	2.85
4	diclofeno	331	1.80	C09AA02	351	1.91	D08A	515	2.81
5	tao	288	1.57	N02BB02	348	1.90	C10A	505	2.75
6	omeparazol	268	1.46	C03CA01	316	1.72	C09A	501	2.73
7	fyrosemida	226	1.23	C01EB16	316	1.72	J01C	435	2.37
8	optobite	222	1.21	J01	311	1.70	N05B	419	2.28
9	ibuprfeno	219	1.19	B01	311	1.70	A10B	394	2.15
10	nolottl	177	0.97	B03BA01	276	1.51	A02B	372	2.03
11	atrovent	175	0.95	D08AL30	275	1.50	C01E	357	1.95
12	metformian	173	0.94	A02BC01	274	1.49	A10	357	1.95
13	auqacel	170	0.93	J07BB	257	1.40	C03C	354	1.93
14	clxane	170	0.93	J01CR02	225	1.23	N06A	354	1.93

Table 13: 10 most frequent values for Farmaco nodes after preprocessing

				Level 6			Level 2		
	Name	Count	%	Name	Count	%	Name	Count	%
0	hta	575	3.69	I10	739	4.74	I10	739	4.74
1	pies	303	1.94	E14	440	2.82	T14	602	3.86
2	dm	269	1.73	J00	351	2.25	C00-C75	488	3.13
3	atdon	226	1.45	I64	310	1.99	E14	459	2.94
4	epoc	183	1.17	Z76.8	302	1.94	J00	351	2.25
5	feril	178	1.14	F17.1	263	1.69	I64	310	1.99
6	iam	165	1.06	Z74	258	1.65	Z76	302	1.94
7	fumadora	163	1.05	E11	252	1.62	M50-M54	301	1.93
8	avc	161	1.03	T14.1	249	1.60	J44	291	1.87
9	anemias	160	1.03	J44.9	237	1.52	F17	265	1.70
10	itu	144	0.92	I46.9	226	1.45	I21	261	1.67
11	defuncio	136	0.87	H26.9	225	1.44	I25	259	1.66
12	cva	134	0.86	I48	217	1.39	Z74	258	1.65
13	dm2	115	0.74	I21.9	205	1.31	E11	253	1.62
14	acxfa	112	0.72	N39.0	201	1.29	E78	249	1.60

Table 14: 10 most frequent values for Diagnostico nodes after preprocessing

					Level 2		
	Name	Count	%		Name	Count	%
0	doloros	987	4.13	0	ciap2:A29	2126	8.91
1	mvc	637	2.67	1	ciap2:A01	1725	7.23
2	tos	603	2.53	2	ciap2:K29	1346	5.64
3	disnea	499	2.09	3	ciap2:R29	1145	4.80
4	fibre	377	1.58	4	ciap2:N29	975	4.08
5	beg	302	1.26	5	ciap2:A03	967	4.05
6	vomito	294	1.23	6	ciap2:R02	864	3.62
7	fetge	280	1.17	7	ciap2:R05	844	3.54
8	feril	231	0.97	8	ciap2:D29	690	2.89
9	diarreas	224	0.94	9	ciap2:R04	686	2.87
10	oplos	170	0.71	10	ciap2:S29	546	2.29
11	dolor toracicco	167	0.70	11	ciap2:L29	494	2.07
12	nauceas	160	0.67	12	ciap2:S19	416	1.74
13	eupneica	144	0.60	13	ciap2:K05	360	1.51
14	nh	140	0.59	14	ciap2:A04	315	1.32

Table 15: 10 most frequent values for SignoSintoma nodes after preprocessing

	Name	Count	%
0	aabdominal	381	2.64
1	eeeii	380	2.63
2	andomen	291	2.01
3	torarcico	264	1.83
4	c lumbar	234	1.62
5	pulmon	214	1.48
6	abd	188	1.30
7	pies	181	1.25
8	colorn	174	1.20
9	torso	169	1.17
10	farige	155	1.07
11	orofaringe	133	0.92
12	peni	109	0.75
13	vertebral	109	0.75
14	cafi	108	0.75

Table 16: 10 most frequent values for ParteCuerpo nodes after preprocessing

Although the preprocessing steps, we can observe that the percentages of appearances are small. This can make us think we will obtain very small supports for the associations. We observe that when generalizing, some codes they barely change, but in other cases like the *Diagnostico* code *C00-C75* which refers to *"Neoplasias malignas"* has a big increase in appearance. So the generalizations will help us obtain different rules.

4.6 Our temporal association rules formulation

As it was mentioned in Section 3.1, during the evolution of this thesis we performed several iterations in the typical phases of a data mining project, analyzing different approaches. In this section, we will present the formulation for computing the temporal rules that we finally used.

Let $\mathcal{J} = I_1, I_2, ...I_m$ be a set of *items* and $\mathcal{D} = S_1, S_2, ...S_n$ be a set of transactions.

Let

$$g^p_{S_i} : \mathcal{P}(\mathcal{J}) \rightarrow \{0,1\}, \quad g^p_{S_i}(X) = \begin{cases} 1 & \text{if } S_i \text{ exists and contains } X \text{ for patient } p \\ 0 & \text{else} \end{cases} \tag{14}$$

Where S_i is visit at time (day) i.

Then

$$g^p_{S_i,t} : \mathcal{P}(\mathcal{J})^2 \rightarrow \{0,1\}, \qquad g^p_{S_i,t}(X,Y) = \min\left(g^p_{S_{i-t}}(X), g^p_{S_i}(Y)\right) \tag{15}$$

Where p is a patient, S_i is a visit at time i, and t the temporal gap in days. This function returns 1 only when patient p has X at time $i-t$ and the same patient p has Y at time i, 0 otherwise. If X or Y is equal to $\emptyset$, in our framework means the existance of a visit at the given time. For example, $g^p_{S_i,t}(X,\emptyset)$ returns 1 if S_{i-t} exists and contains X, and S_i exists for a given patient p. Having $T = [t_1, t_2]$ $(t_1 < t_2)$, in our case, t_1 will be *min_gap* and t_2 *max_gap*.

$$h_T : \mathcal{P}(\mathcal{J})^2 \rightarrow \mathbb{N}, \qquad h_T(X,Y) = \sum_p \sum_{S_i \in D} min(\sum_{t \in T} g^p_{S_i,t}(X,Y), 1) \tag{16}$$

h_T is the count support for X, Y. The inner sum gives, for a given patient p and set of visits S_i, how many times has X at time $i-t$ and Y at time i. The min function ensures that if the association $X \overset{T}{\Rightarrow} Y$ exists, we only count it once per each consequent. The $\sum_{S_i \in D}$ sum is done for all visit sets, and the outer sum is done for all the patients. According to these formulas this, we define the following indicators:

$$\textit{support } (X \overset{T}{\Rightarrow} Y) \;\; = \frac{h_T(X,Y)}{h_T(\emptyset,\emptyset)} \tag{17}$$

$$\textit{confidence } (X \overset{T}{\Rightarrow} Y) \;\; = \frac{h_T(X,Y)}{h_T(X,\emptyset)} \tag{18}$$

$$\mathit{rulesRespectConsequent}\ (X \overset{T}{\Rightarrow} Y) \ = \frac{h_T(X,Y)}{h_T(\emptyset,Y)} \tag{19}$$

$$\mathit{lift}\ (X \overset{T}{\Rightarrow} Y) \ = \frac{\mathit{support}(X,Y)}{\mathit{support}(X,\emptyset) * \mathit{support}(\emptyset,Y)} \tag{20}$$

$$\mathit{leverage}\ (X \overset{T}{\Rightarrow} Y) \ = \mathit{support}(X,Y) - \mathit{support}(X,\emptyset) * \mathit{support}(\emptyset,Y) \tag{21}$$

The *rulesRespectConsequent* indicator was asked by the doctors during the evaluation phase and will be used in future sections.

4.7 Temporal association rules algorithm

Using the formulation defined in the previous section, we will present the algorithm we used to extract the rules. First, we will define the principal data structures that we will use, and afterward, we will present the pseudocode of the algorithm used.

4.7.1 Principal data structures

We will work with two main data structures. The first one is a table that contains, for each patient and visit dates, in one column a list of all the *Diagnostico* instances contained in that visit for that patient. We keep the *Diagnostico* instances because they are the ones that we want to have in the RHS of the extracted rules, but this can be parametrized in case we wanted to obtain rules with other types of nodes in the consequent. The second column contains all the possible nodes that can be counted as antecedents: all the nodes that belong to the same patient and the date of the visit fulfill the *min_gap* and *max_gap* restrictions. Table 18 shows a simple example of the data in Table 17. In this example, we will assume all transactions are visits from the same patient and all the items are of type *Diagnostico*. We will use the *ID* of each row as the date of the visit. We will take a *min_gap* = 1 and *max_gap* = 2. In our real problem, each item is composed of all the attributes mentioned in Section 4.1.1, and the visits obviously do not have to be consecutive.

ID	Transactions
1	A, C, E
2	B, D
3	B, C
4	A, B, C, D
5	A, B
6	B, C
7	A, B
8	A, B, C, E
9	A, B, C
10	A, C, E

Table 17: Original data

$\xRightarrow{(1,2)}$

ID	Antecedent	Consequent
1	-	A, C, E
2	A, C, E	B, D
3	A, C, E, B, D	B, C
4	B, D, B, C	A, B, C, D
5	B, C, A, B, C, D	A, B
6	A, B, C, D, A, B	B, C
7	A, B, B, C	A, B
8	B, C, A, B	A, B, C, E
9	A, B, A, B, C, E	A, B, C
10	A, B, C, E, A, B, C	A, C, E

Table 18: Temporal data structure

The second data structure is a binary version of the last table. This table represents, for each row of the last table, if a given variable is in the antecedent and consequent. In this case, the item *A'* in Table 19 represents the variable we want to analyze of the item *A* of Table 18. For example, if we are searching association rules based on the code of the item, *A'* will be a code.

ID	A', Ant	B', Ant	C', Ant	D', Ant	E', Ant	A', Con	B', Con	C', Con	D', Con	E', Con
1	0	0	0	0	0	1	0	1	0	1
2	1	0	1	0	1	0	1	0	1	0
3	1	1	1	1	1	0	1	1	0	0
4	0	1	1	1	0	1	1	1	1	0
5	1	1	1	1	0	1	1	0	0	0
6	1	1	1	1	0	0	1	1	0	0
7	1	1	1	0	0	1	1	0	0	0
8	1	1	1	0	0	1	1	1	0	1
9	1	1	1	0	1	1	1	1	0	0
10	1	1	1	0	1	1	0	1	0	1

Table 19: Binary data structure

4.7.2 Temporal association rules extraction

Using the formulation previously described, we used Algorithm 1 to extract the temporal association rules. The presented algorithm is a modification of the one explained in Section 2.2. The algorithm will use the principal data structures described in Section 4.7.1. We must remember that to build those tables we needed the *min_gap* and *max_gap* variables. Apart from these requirements, the algorithm takes as input the following variables:

- *min_sup:* The minimum accepted support.
- *min_conf:* The minimum accepted confidence.
- *attribute:* The attribute where we want to find the association rules. This can take values *code* or *raw_text*.
- *max_antecedents:* The maximum number of antecedents accepted in the rule.
- *valid_cons:* The consequents we want to have in the RHS of the rule. We will do several executions with several parametrizations, this will allow us to filter some consequent if needed.

We also want to define what we called the *Antecedents of Interest (AoI)* and *Consequents of Interest (CoI)*. The *AoI* is composed of all those antecedents whose support is greater than the *min_sup*. The *CoI* is composed of all those consequents in *valid_cons* whose support is greater than the *min_sup*.

We developed a very parametrizable code, not only regarding the *support* and *confidence* indicators but allowing the extraction rules with different time gaps, amount and types of antecedents, types of consequents, using different levels of abstraction on each concept and allowing to search associations with different attributes of the data. These will be very useful because when performing the evaluation we will need to adapt the results to the requirements of the doctors.

Algorithm 1 Temporal Rules Extraction

```
$Rules \leftarrow \emptyset$
$A \leftarrow \emptyset$
for all $a_i \in CoI$ do
    $cons \leftarrow a_i$
    for all $a_j \in AoI$ do
        if $\text{Sup}(\{a_j\} \overset{T}{\Rightarrow} \text{cons}) \geq min\_sup$ and $\text{Conf}(\{a_j\} \overset{T}{\Rightarrow} \text{cons}) \geq min\_conf$ then
            Compute $\{a_j\} \overset{T}{\Rightarrow} cons$ time metrics
            $Rules \leftarrow Rules \cup \{(\{a_j\} \overset{T}{\Rightarrow} cons)\}$
            $A(cons) \leftarrow A(cons) \cup \{a_j\}$
        end if
    end for
end for
$k \leftarrow 2$
repeat
    $nextFeasibleRules \leftarrow getNewFeasibleRules(A, k)$
    for all $r \in nextFeasibleRules$ do
        if $\text{Sup(r)} \geq min\_sup$ and $\text{Conf(r)} \geq min\_conf$ then
            Compute $r$ time metrics
            $Rules \leftarrow Rules \cup \{r\}$
            $A(cons) \leftarrow A(cons) \cup r_{antecedent}$
        end if
    end for
until $nextFeasibleRules$ is empty or $k \geq max\_antecedents$
```

The *Sup* and *Conf* functions, that compute the support and confidence of a rule, can be fastly computed using the formulations in Section 4.6 and the binary data structure. Regarding the *getNewFeasibleRues(A, k)* function, it computes the new feasible rules with k antecedents, based on the idea that if an itemset is not frequent, any of its superset will never be frequent. Those possible rules will need to be checked if they fullfill the *min_suport* and *min_conf* conditions.

For each rule that fullfils the *min_sup* and *min_conf* conditions, we will compute the time metrics. Let $ED(i_1, i_2)$, abreviation from *elapsed days*, be a function that computes the days between two items i_1 and i_2.

$$ED(i_1, i_2) \quad = |i_{2_{date}} - i_{1_{date}}| \tag{22}$$

Then, we will compute the time between antecedents (TA), and the time between antecedents an consequent (TC).

$$TA(\{a\} \Rightarrow c) \quad = \sum^{|rules|} \frac{\frac{2! * (|a|-2)! * \sum_i^{|a|} \sum_{j>i}^{|a|} ED(a_i, a_j)}{|a|!}}{|rules|} \tag{23}$$

$$TC(\{a\} \Rightarrow c) \quad = \sum^{|rules|} \frac{\frac{\sum_i^{|a|} ED(a_i, c)}{|a|}}{|rules|} \tag{24}$$

Basically in Equation 23 and 24 we are computing an average of averages. In Equation 23, we are averaging the time elapsed between all pairs of instances included in the antecedent of the rule. The factorial terms are obtained when computing the number of two length combinations without repetitions that can be obtained from the antecedent. We also average the resulting value with all the times the rule is fulfilled. A similar idea is applied in Equation 24, where we compute the average time elapsed between all antecedents and the consequent and average the resulting value for all the times the rule is fulfilled.

5 Evaluation

In this section, we will first explain how the evaluation was performed, and afterward, we will present the results obtained.

5.1 Description of the evaluation phase

As described in Section 3.1, the evaluation phase consisted of two steps: A self-evaluation step, and an external evaluation step. The external evaluation step is especially important because it has been designed not only for this thesis but will be the process of evaluation of the entire TADIA-MED project.

During the self-evaluation step, we used our algorithm with synthetic data as it could be the one shown in Table 17. We discussed the results comparing them with the previously obtained using other formulations, to be sure that our formulation reflected what we intended. After several iterations of this process, once we were sure the formulation was correct, we moved to the external evaluation step.

In the external evaluation step, we presented the obtained rules to the doctors of the IDIAP Jordi Gol. Before start correcting the codes, several meetings were performed. Those meetings were attended by Pere Torán, Concepció Violan, Víctor Miguel López, David Lacasta, and Luis Rodriguez on behalf of the IDIAP Jordi Gol, and Alicia Ageno, Neus Català, Jordi Turmo, Lluís Padró, and the author of this thesis Oriol Borrell, on behalf of the UPC.

The first purpose of the meetings was to describe to the doctors the format and the possible interpretation of the rules that we were extracting. The second main reason for the meetings was to reach a consensus in the format of evaluation. We agreed that for each rule the *correctness* and *relevancy* of the rule would be evaluated. Regarding *correctness* of a rule we agreed on the following labels:

- **Totally incorrect:** The rule is completely wrong. For abbreviation purposes, when presenting the results we will call this label *"Incorrect"*.
- **Totally correct:** The rule is completely right. For abbreviation purposes, when presenting the results we will call this label *"Correct"*.
- **Partially correct (for the temporal aspect):** The rule is correct, but the time measures are not the expected ones. For abbreviation purposes, when presenting the results we will call this label *"Par-Temporal"*.
- **Partially correct (for the clinical aspect):** The rule is partially correct because some antecedent is missing or leftover. For abbreviation purposes, when presenting the results we will call this label *"Par-Clinical"*.

- **Partially correct (for both aspects):** The rule is partially correct because some antecedent is missing or leftover and the time measures are not the expected ones. For abbreviation purposes, when presenting the results we will call this label *"Par-Both"*.

Regarding the *relevancy* of a rule, the doctors must classify each rule in the following groups:

- **Not relevant:** The rule has no significant clinical relevance. For abbreviation purposes, when presenting the results we will call this label *"No-Rel"*.
- **Relevant and known:** The rule has significant clinical relevance and is a well-known rule. For abbreviation purposes, when presenting the results we will call this label *"Rel-Known"*.
- **Relevant and unknown:** The rule has significant clinical relevance and is a rule that must be studied because is not a known rule. For abbreviation purposes, when presenting the results we will call this label *"Rel-Unknown"*.

Finally, we asked for a brief *justification* of the answers. We also agreed that the external evaluation phase would consist of two steps:

1. An excel document will be provided by the author of this thesis to the three doctors that will correct the rules. These doctors are: Víctor Miguel López, David Lacasta, and Luis Rodriguez. For each rule we will include the following information: Codes of the LHS, LHS, Code of the RHS, RHS, the *rulesRespectConsequent* indicator, TA and TC. This initial correction will be performed by each doctor individually. Each doctor will provide to the author of the thesis the individual corrections.
2. The author will merge the three individual corrections into an online document, that will be sent to the doctors. The three doctors will have to agree on a unique answer for each rule.

To decide the explained evaluation process, several iterations were made. We performed two test evaluations with 20 rules. The intention was to consolidate the evaluation criteria, familiarize the doctors with the results and their interpretation. We also wanted to detect the information that was useful for the doctors.

During the first iteration, we found that there were not a lot of coincidences on the answers of the individual phase. We detected that this was caused for two reasons: The meaning of the labels was not clear enough, and not all the doctors were understanding the same diagnosis explained in the rules. To solve the first problem we added the label *Partially correct (for both aspects)* in the *correctness*, and clarified the meaning of each label. For the second problem, we detected that the problem was in the *cim10* codes that start with *Z* because they contain very disperse diagnoses. We detected the codes that may be confusing and we decided to manually correct the code description based on the *raw_text* attributes.

During these revisions we also detected that neither the *confidence* nor *support* of a rule was useful for the doctors. They wanted to know, out of all the cases that suffer the consequent, how many medical cases had the antecedents in the elapsed gap. That is why we defined the *rulesRespectConsequent* indicator defined in Section 4.6.

The last problem we detected during the test iterations was that is very different, in terms of relevancy and temporality, to have in the RHS *refredat* (cold, in English) than to have *PADES*. PADES is a palliative care program in Catalonia. If we have a cold, it is interesting to know what happened during a gap of 3-15 days, but in case of PADES, we would like to know what happened in the last 30-300 days. Furthermore, the PADES diagnose is much more relevant than the cold. For this reason, with the help of the doctors, we classified the diagnoses into three groups: short (5 to 30 days), medium (10 to 120 days), and long (30 to 300 days) term. We also asked each doctor to specify the most important diagnoses, among the most frequent ones in the dataset. The important diagnoses and the gaps proposed per each doctor can be observed in Table 28 in Appendix A.

It was previously agreed that the IDIAP doctors would correct 10 packs of 30 rules. The packs intend to deliver rules with different parameterizations. Despite all the rules were already computed from our part, we decided to present only two packs, to get the corrections on time and include them in this thesis. The rest of the rules will be sent to the doctors after the submission of this report. In the next subsection, we will analyze the results of the evaluation of these first two packs of rules.

5.2 Results

We presented two packs to the doctors. The first one was focused on finding short-term time rules (clinical situations that occur within 5 to 30 days). The second pack was focused on medium-term time rules (clinical situations that occur within 10 to 120 days). With our executions we can obtain thousands of rules, depending of the *min_sup*, *min_conf* and *max_antecedents* parameters. To select the rules that would be presented to the doctors, for each *min_gap* - *max_gap* parametrization we followed the next steps:

1. Execute our algorithm with a very small *min_sup* and *min_conf* parameters, to obtain a lot of rules.

2. Order the rules by consequent.

3. For each consequent, if a big number of rules were obtained, using a post-filter, increase the *min_sup*. Select the rules using the *lift* indicator.

We decided to perform these steps because in these first two packs we wanted to ensure that we had a variety of consequents. These first and second packs are computed from the 23X2 preprocessed *csv*. The first and second packs can be observed in Table 29 and Table 30 in Appendix B and Appendix C. The corrections performed by the doctors can be observed in Table 31 and Table 32 in Appendix D and Appendix E.

The first pack is composed of 25 rules. Table 21 shows a summary of the obtained results during the different external evaluation phases. The first thing we can observe looking at the table is that it does not exist a rule classified with the label *Par-Temporal* or *Par-Both* in the *Correctness*. We can interpret that we are parameterizing correctly the *min_gap* and *max_gap* parameters in this pack.

The other easy thing to observe in Table 21 is that no rules were classified with the *Rel-Unknown* label for the *Relevancy*. This was quite surprising because during the test iteration we obtained several results with that label. We must take into account that we are using small *min_gap* and *max_gap* parameters and it's easy for the doctors to track what happens in a small period of time. We expect to obtain more *Rel-Unknown* labels when correcting the long-term rules. For the individual phase of this execution, we can observe that the majority of the rules were corrected with *Correct* or *Par-Clinical* for the *Correctness* and *Rel-Known* for the *Relevancy*. This is confirmed after the second phase of evaluation (the agreement phase), were we can observe that no *Incorrect* values were obtained, and the *Rel-Known* is the most common result for the *Relevancy*. We don't interpret this as a bad result, because it means that we are finding well-known associations. We need to observe what happens with different parametrizations to extract additional conclusions.

	Incorrect	Correct	Par-Temporal	Par-Clinical	Par-Both	No-Rel	Rel-Known	Rel-Unknown
Victor	2	18	0	5	0	3	22	0
David	2	17	0	6	0	3	22	0
Luis	4	11	0	10	0	7	18	0
Agreement	0	16	0	9	0	2	23	0

Table 21: Summary of the results, pack 1

We can observe in Table 23 that the 3 doctors answered the same *Correctness* label in only 9 rules. Regarding the *Relevancy*, we reached higher consensus.

		Correctness	Relevancy
David	Victor	11	19
David	Luis	14	19
Victor	Luis	12	17
3 Doctors		9	15

Table 23: Agreement between doctors in pack 1

From this first pack, we want to highlight the rule showed in Equation 25 (rule 3, pack 1). The number above the arrow represents the TC parameter. Despite this rule was mostly annotated with the *Par-Clinical* and *No-Rel* labels, the doctors mention that if the A.A.S (acetylsalicylic acid) were included as an antipyretic the rule would be true. But if A.A.S were not included, the association would be relevant and surprising. Therefore, we think this rule must be analyzed using a lower level of generalization.

$$\boxed{\text{Otros analgésicos y antipiréticos} \overset{15,6}{\Longrightarrow} \text{Infarto agudo del miocardio}} \tag{25}$$

Other results are not very interpretable. Equation 26 (rule 4, pack 1) has very disperse classification in both *Correctness* and *Relevancy* aspects. Looking at the justifications, Luis does not find any relation between LHS and RHS, David on the contrary mentions that there is a clear relation, and Victor states that is the habitual procedure. Finally, the three doctors agreed that this was the habitual procedure, and they classified the rule as *Correct* and *Rel-Known*. This a clear example of the complexity of the evaluation process.

$$\boxed{\begin{array}{lcl} \text{Antisépticos y desinfectantes,} & & \\ \text{Dolor generalizado/múltiple,} & \overset{13,7}{\Longrightarrow} & \begin{array}{l}\text{Traumatismo de regiones} \\ \text{no especificadas del cuerpo}\end{array} \\ \text{Cambios en el color de la piel} & & \end{array}} \tag{26}$$

The second pack is composed of 34 rules. Table 25 shows a summary of the obtained results during the different external evaluation phases. Looking at the mentioned table we can extract similar conclusions to the ones that we mentioned from the first pack. *Correct* and *Rel-Known* are the most frequent labels obtained. But in this case, we have more appearances of *Incorrect* and *No-Rel*. These results must be taken into account when preparing the future packs.

	Incorrect	Correct	Par-Temporal	Par-Clinical	Par-Both	No-Rel	Rel-Known	Rel-Unknown
Victor	5	15	0	2	3	6	19	0
David	9	11	4	1	0	11	14	0
Luis	5	11	0	8	1	5	19	1
Agreement	7	21	1	3	2	8	26	0

Table 25: Summary of the results, pack 2

Looking at Table 27, we can observe that in the second pack we obtained fewer similarities in the results of the individual evaluation phase. We must remember that this second pack contains 9 more rules than the first one. But if we deeply analyze how these differences are distributed, we observe that most of the rules have at least two equal labels per *Correctness* and *Relevancy*. We expect the doctors to reach an easy agreement on the answers.

		Correctness	Relevancy
David	Victor	19	24
David	Luis	12	22
Victor	Luis	13	22
3 Doctors		8	17

Table 27: Similarities between doctors in pack 2

We want to mention a rule where no agreement was reached concerning the *Correctnes*. In the rule represented in Equation 27 (rule 3, pack 2), Victor responded *Correct*, David responded *Par-Temporal* and Luis responded *Incorrect*. It would be interesting to analyze with the doctors the reasons for each answer because it could be that the RHS is too unspecific. However, we want to highlight that when there is a LHS or RHS that contains "Otros (qualquier cosa)", in English "Others (whatever)", the doctors have the CIM10 codes at their disposal. Then, they can check which are the rest of codes of the same type (at the same level in the CIM codes hierarchy), and rule them out.

$$\boxed{\begin{array}{lcl}\text{Otros analgésicos y antipiréticos,} & \overset{69{,}5}{\Longrightarrow} & \text{Otros trastornos de los} \\ \text{Dolor generalizado/múltiple} & & \text{tejidos blandos}\end{array}} \tag{27}$$

In this second pack, we also found some typical symptomatology of a diagnosis. This can be the case of the rule represented in Equation 28 (rule 16, pack 2). In this case, all the doctors coincide in classifying this rule as *Correct* and *Rel-Known*. All the doctors justify the result saying that all the instances in the LHS are signs of the RHS.

$$\boxed{\begin{array}{l}\text{Otros signos/síntomas del aparato respiratorio,}\\ \text{Otros analgésicos y antipiréticos,}\\ \text{Tos,}\\ \text{Dolor generalizado/múltiple}\end{array} \quad \overset{59,0}{\Longrightarrow} \quad \text{Neoplasia maligna}} \tag{28}$$

Finally, we want to mention the rule where the *Rel-Unknown* label for the *Relevancy* was obtained. This rule is the one shown in Equation 29 (rule 23, pack 2). For this rule its clear that *Tabaco* causes *Neoplasia maligna*, but it's surprising that *Trastornos mentales y del comportamiento debidos al uso de tabaco* provokes *Neoplasia maligna*. Finally the doctors agreed that this was a *Rel-Known* rule, but we think that this rule must be evaluated using a lower level of abstraction.

$$\boxed{\begin{array}{l}\text{Trastornos mentales y del}\\ \text{comportamiento debidos al uso de tabaco}\end{array} \quad \overset{59,2}{\Longrightarrow} \quad \text{Neoplasia maligna}} \tag{29}$$

In this section, we highlighted some results. To increase the readability of this thesis, for each described rule, we only indicated the antecedents, consequent, and the *TC* parameter. As we previously mentioned, all the rules can be observed in Table 29 and Table 30 in Appendix B and Appendix C, and the annotations performed by the doctors can be observed in Table 31 and Table 32 in Appendix D and Appendix E. We are aware that the support obtained for the rules was very small. This is caused for two reasons: First, because the short-term and medium-term time codes are the less frequent ones in our dataset. This can be observed in Table 28 in Appendix A. Slightly better results were obtained in the long-term time rules. The second reason is that our dataset is only composed of visits of only 320 patients. We are aware that is difficult to extract relevant conclusions with such a small sample. But, our main objective was to validate our methodology, and observing that the associations obtained are mostly correct in terms of clinical content and chronology, we consider our goal as achieved.

6 Conclusions

After performing all the proposed tasks over the dataset, we can conclude we accomplished all the goals defined at the beginning of the project. We were able to extract temporal association rules on multimorbid patients, allowing us to express the medical instances in different levels of abstraction.

First, an extensive analysis of the literature was performed. We searched for different ways to face our problem, only a subset of the reviewed documentation is detailed in the thesis. The state-of-the-art research gave us a basic idea of where to start our project and its possibilities.

Second, a detailed description of the data was performed and the errors on the data found were presented. Despite we are not the only ones working with these data, no extensive documentation of the data existed. This made the data understanding process more difficult. The description performed in this thesis will be very useful for new people that may work in the TADIA-MED project. It will also be very useful for the researchers that are working on the codification of new data.

Third, a solid formulation with a very parametrizable temporal association rules algorithm was presented. We attached a lot of importance to making our algorithm very parametrizable, not only regarding the *support* and *confidence* indicators but allowing the extraction rules with different time gaps, amount and types of antecedents, types of consequents, using different levels of abstraction on each concept and allowing to search associations with different attributes of the data. We discussed the best parametrization. Using those values we were able to extract temporal association rules on multimorbid patients. Our work is intended to be a solid basis to keep evolving into a powerful tool for primary health care rule extraction. Moreover, given the flexibility of our method, it might be easily extended to be applied to the extraction of temporal association rules in other domains.

Fourth, an evaluation process for those rules was defined and initialized. This was not a trivial part because it was not that easy to clarify to the doctors the interpretation that could be made for each rule and how the rules needed to be evaluated. We spent a significant effort on this part. We thought it was necessary to correctly define this evaluation part, as it will be the one used in the correction of future projects developed with these data.

Therefore, all the required tasks were performed successfully, providing me with a lot of information and insights from the dataset. It was also very interesting to work with this dataset because I learned to work with a more real-world-oriented project, facing the difficulty of working in a context that is outside my field of expertise as the primary health care context. It was a very interesting project, I enjoyed a lot and I learned many interesting concepts I will surely be using in my future professional career. The realization of this thesis and the completion of the Master in Innovation and Research in Informatics majored in Data Science, presented me a very interesting job opportunity in an external company that I will be very proud to accept after the confirmation of my graduation.

6.1 Future Work

As mentioned before, several tasks must be performed before the finalization of the project. The first important task is to select the next eight packs of rules that will be sent to the doctors. We should use the already received feedback for selecting the remaining eight packs. Once obtained the results of all the packs, global conclusions of the results should be obtained to give this part of the project as approved.

We also mentioned that our work is intented to be a solid base that should be extended in future projects. One possible extension would be the possibility of finding rules of the form:

$$\{a_1\} \overset{T_1}{\Longrightarrow} \{a_2\} \overset{T_2}{\Longrightarrow} \cdots \overset{T_{n-1}}{\Longrightarrow} \{a_n\} \overset{T_n}{\Longrightarrow} \text{c} \tag{30}$$

Where $\{a_i\}$ are sets of clinical instances that happen during the same visit, and c is a diagnose. This would increase the notion of the time elapsed between antecedents, and more importantly, the order in which the antecedents appear. One of the potential drawbacks of this proposal is that it will probably provoke a significant decrease on the *support* of the rules, but this must be studied.

Another interesting aspect that can be done is to increase the years included in the study. In our thesis, we studied multimorbid patients that were visited between 2010 and 2016. Increasing the years of the study will provoke an increase in the number of patients, which will lead to more solid results. This must be performed carefully, because of the appearances of new diagnoses like *COVID-19*.

Also, several extensions can be applied regarding the analysis of the *raw_text* attributes. Although our algorithm can be executed for extracting rules on this attribute, we focused on obtaining temporal association rules based on the *code* attribute. Moreover, the generalization we performed was also based on this attribute, but using the *raw_text* attribute, some hierarchies of concepts, text embeddings, or other NPL techniques could be applied in case of collecting more data.

Finally, other different approaches must be studied. As we mentioned our data was represented as a graph. We have simply extracted the data from Neo4J, but we have not taken advantage of the graph structure itself due to lack of time. Several other extensions of the apriori algorithm based on graphs are presented in the book *"Top 10 algorithms in data mining"* [19]. Some interesting extensions could be the gSpan algorithm [20] or the framework for mining frequent subgraphs from labeled graphs presented by Akihiro Inokuchi [10].

References

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. Citeseer. 1994, pp. 487–499.

[2] Rakesh Agrawal and Ramakrishnan Srikant. "Mining sequential patterns". In: *Proceedings of the eleventh international conference on data engineering*. IEEE. 1995, pp. 3–14.

[3] James F Allen. "Towards a general theory of action and time". In: *Artificial intelligence* 23.2 (1984), pp. 123–154.

[4] *Anatomical Therapeutic Chemical (ATC) Classifications*. URL: `https://www.who.int/tools/atc-ddd-toolkit/atc-classification`. (accessed: 28.05.2021).

[5] *Beth Israel Deaconess Medical Center*. URL: `https://www.bidmc.org/`. (accessed: 13.06.2021).

[6] *Clasificación Internacional de Enfermedades*. URL: `https://icdcode.info/espanol/cie-10/codigos.html`. (accessed: 28.05.2021).

[7] S Concaro et al. "Temporal data mining for the analysis of administrative healthcare data". In: *IDAMAP Workshop, Washington*. 2008.

[8] *Cypher Query Language*. URL: `https://neo4j.com/developer/cypher/`. (accessed: 15.06.2021).

[9] *IDIAP Jordi Gol*. URL: `https://www.idiapjgol.org/index.php/en/`. (accessed: 13.05.2021).

[10] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. "A general framework for mining frequent subgraphs from labeled graphs". In: *Fundamenta Informaticae* 66.1-2 (2005), pp. 53–82.

[11] *Lista de Códigos CIE10 en español como librería python*. URL: `https://pypi.org/project/cie/`. (accessed: 30.05.2021).

[12] *MyDisk*. URL: `https://rdlab.cs.upc.edu/mydisk/`. (accessed: 15.06.2021).

[13] *Neo4j*. URL: `https://neo4j.com/`. (accessed: 15.06.2021).

[14] *Python*. URL: `https://www.python.org/`. (accessed: 15.06.2021).

[15] Lucia Sacchi et al. "Data mining with temporal abstractions: learning rules from time series". In: *Data Mining and Knowledge Discovery* 15.2 (2007), pp. 217–247.

[16] *Systematized Nomenclature of Medicin*. URL: `https://www.snomed.org/`. (accessed: 28.05.2021).

[17] *TADIA-MED*. URL: `https://futur.upc.edu/28881334`. (accessed: 13.05.2021).

[18] *TALP*. URL: `http://www.talp.upc.edu`. (accessed: 13.05.2021).

[19] Xindong Wu and Vipin Kumar. *The Top Ten Algorithms in Data Mining*. 1st. Chapman & Hall/CRC, 2009. ISBN: 1420089641.

[20] Xifeng Yan and Jiawei Han. “gspan: Graph-based substructure pattern mining”. In: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE. 2002, pp. 721–724.

[21] Li Zhan, Fusheng Yu, and Huixin Zhang. “A fast algorithm for mining temporal association rules based on a new definition”. In: *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).* IEEE. 2017, pp. 1548–1553.

A Time intervals and relevance of diagnoses

The following table shows, the gap proposed and importance per each doctor. *T* stands for *Time*, and *I* for *Importance*. Regarding the letters inside the brakets, *L* refers to *Luis*, *D* to *David* and *V* to *Victor*. *ID* is the *cim10* code of the diagnose. # represents the number of times the code appears in the original dataset.

#	ID	TEXT	T (L)	T (V)	I(L)	I(D)	I(V)	#I
739	I10	HTA	120-300	30-300	1	1	1	3
440	E14	diabetis	120-300	60-300	1	1	1	3
351	J00	refredat	3-15	3-15	0	0	0	0
309	I64	ICT	30-300	30-300	1	0	1	2
302	Z76.8	PADES	30-300	30-300	1	0	1	2
263	F17.1	fumador	120-300	15-300	1	0	0	1
258	Z74	atdom	30-120	15-300	1	0	1	2
252	E11	dm2	30-300	60-300	1	1	1	3
249	T14.1	ferida	3-20	3-20	0	0	0	0
237	J44.9	MPOC estable	30-300	30-300	1	0	1	2
225	H26.9	Catarata a ull esquerre	60-120	60-120	0	0	0	0
223	I46.9	defunció	30-300	30-300	0	0	1	1
217	I48	ACXFA	30-300	15-60	1	1	1	3
205	I21.9	IAM	30-300	15-30	1	1	1	3
201	N39.0	ITU	3-15	3-15	0	0	0	0
200	T14.2	Fractura de 5º MTT	3-15	3-15	0	0	0	0
197	E78.9	DLP	30-300	60-300	1	0	1	2
194	D64.9	anemias	30-120	30-120	0	1	1	2
180	I25.9	Insuficiència cardíaca	60-300	60-300	0	1	1	2
137	C18.9	neopliasia	30-300	30-180	1	1	1	3
132	W19.9	Caiguda	10-120	10-120	0	0	0	0
126	F06.7	deterioro congnitivo	30-300	30-300	1	1	1	3
120	M54.5	Lumbalgia	30-300	30-300	0	0	0	0
113	E16.2	hipoglucèmies	5-20	5-20	0	0	1	1
112	C34.9	tumor pulmonar	30-300	30-180	1	1	1	3
106	I63.9	AVC isquemico	5-120	1-30	1	1	1	3
105	I50.9	IC	30-300	30-300	1	1	1	3
100	H36.0	retinopatia diabetica	30-300	180-300	1	1	1	3
99	R51	cefalea	5-120	5-120	0	0	0	0
97	I51.7	cardiomegalia	30-300	30-300	0	1	0	1
90	J22	sobreinfección respiratoria	10-60	10-60	0	0	1	1
90	N18.9	I.Renal	120-300	120-300	0	1	1	2
89	H61.2	Tapon de cerumen	30-90	30-90	0	0	0	0
88	G45.9	AIT	30-300	5-300	1	1	1	3
88	J18.9	Pneumonia	5-30	1-15	1	1	1	3

83	R32	INCONTINENCIA URINARIA	30-300	30-300	0	1	0	1
82	E66.9	Obesa	30-300	30-300	0	0	0	0
79	M19.9	ARTROSI	30-300	30-300	0	1	0	1
79	T14.0	Hematoma cerebral cortical frontal	5-60	5-60	0	0	1	1
77	F03	demencia	30-300	90-300	0	1	1	2
74	N40	HBP	30-300	30-300	0	0	0	0
74	T78.4	AL.LERGIA	30-300	30-300	0	0	1	1
73	I80.2	TVP femoropoplitea dreta	30-120	5-15	0	1	1	2
72	H91.9	hipoacusia OD	30-300	30-300	0	0	0	0
72	M17.9	gonartrosis	30-300	30-300	0	1	0	1
70	K46.9	HERNIA INGUINAL IZDA	30-300	30-300	0	0	1	1
69	M25.5	gonalgia derecha	30-300	30-300	0	0	0	0
67	G81.9	paresia	5-120	1-30	0	0	1	1
65	I50.0	INSUFICIÈNCIA CARDÍACA CONGESTIVA	30-300	30-300	0	1	1	2
65	I73.9	CLAUDICACION INTERMITENTE	30-300	30-180	0	1	1	2
63	F41.9	ansietat	5-120	5-120	0	1	1	2
62	G46.7	INFARTOS LACUNARES SUBCORTICALES	30-300	30-300	0	1	1	2
58	J40	proceso bronquial	5-200	5-200	0	0	1	1
58	R42	MAREIG	5-120	5-120	0	0	0	0
57	I20.9	angor d' esforç	30-300	30-180	1	1	1	3
57	M54.2	CERVICALGIA	30-300	30-300	0	0	1	1
57	M77.9	Osteofitos	30-300	30-300	0	0	1	1
55	R63.0	Anorèxia	10-120	10-120	0	0	1	1
53	C20	neoplasia de recto	100-300	60-180	1	1	1	3
53	D50.9	ANEMIA FEROPENICA	10-120	10-120	0	1	1	2
53	G20	parkinsonismo	10-300	10-300	1	1	1	3
53	R39.1	sd miccional	10-120	10-120	0	0	0	0
51	A09	gastroenteritis	5-15	5-15	0	0	0	0
51	M62.4	CONTRACTURA	5-15	5-15	0	0	0	0
50	J44.1	EPOC reagudizada	5-30	5-30	1	0	1	2
49	R73.9	hiperglicemia basal	100-300	100-300	0	0	0	0
48	J20.9	bronquitis aguda	5-30	5-30	0	0	0	0
48	Z63.6	dependència	100-300	100-300	1	0	1	2
47	L89	NAFRA PER DECÚBIT	15-120	15-120	0	0	1	1
47	T81.9	ferida quirúrgica	5-15	5-15	0	0	0	0

46	C44.9	CARCINOMA BASOCELULAR	30-120	30-180	1	0	1	2
46	I20.0	angor de esfuerzo	5-100	30-180	1	1	1	3
46	I84.9	hemorroides	2-30	2-30	0	0	0	0
45	L30.9	eczemas	5-100	5-100	0	0	0	0
45	T14.9	traumatisme peu d	5-30	5-30	0	0	0	0
44	L03.9	CELULITIS	530-90	530-90	0	0	0	0
44	Z73.9	Pacient pluripatologic	30-300	30-300	1	0	1	2
43	R55	CUADRO SINCOPAL	10-100	10-100	0	0	1	1
40	G30.9	enf. Alzheimer moderada	30-300	30-300	1	1	1	3
39	E03.9	HIPOTIROIDISME	30-300	30-300	0	1	1	2
38	R52.9	odontalgia	5-100	5-100	0	0	0	0
36	A16.2	TBC	30-300	30-300	1	0	1	2
36	F05.9	SD confusional	5-15	5-15	1	1	1	3
36	J02.9	Faringitis aguda	5-30	5-30	0	0	0	0
36	T06.0	TCE	5-60	5-60	0	0	1	1
35	I65.2	dues plaques calcificades puntiformes	30-300	30-300	0	0	0	0
35	K44.9	hernia de hiatu	15-100	15-100	0	1	1	2
35	M85.9	Osteopénia	30-300	30-300	1	0	1	2
34	B49	Micosis	5-60	5-60	0	0	0	0
34	M21.7	dismetria entre extremitats	30-300	30-300	0	1	0	1
33	M81.9	Osteoporosis	30-300	60-180	1	1	1	3
33	N20.9	litiasi renal bilateral	30-300	30-300	0	1	1	2
33	Z93.3	colostomia	30-300	30-300	0	1	1	2
32	I34.0	Insuf mitral severa	30-300	30-300	1	1	1	3
31	I80.9	flebitis	10-300	10-300	0	1	0	1
31	K30	DISPÈPSIA	15-300	15-300	0	1	1	2
31	K80	colelitiasis	30-300	60-180	1	1	1	3
31	M16.9	coxastrosis	30-300	30-300	0	1	1	2
31	R00.1	bradicardia ritmica	10-300	10-300	0	0	0	0
31	R10.1	EPIGASTRALGIAS	5-100	5-100	0	0	0	0
30	C61	neoplasia de próstata	30-300	30-300	1	1	1	3
30	F41.2	S, DEPRESIVO DE BASE	30-300	30-300	1	1	0	2
30	M54.3	ciatalgias	30-300	30-300	0	1	0	1
30	N17.9	IRA	10-100	10-100	1	0	0	1
29	F32.9	depressió	15-300	15-300	0	1	1	2
29	H10.9	Conjuntivitis	5-60	5-60	0	0	0	0
29	I99	isquemia	30-300	5-30	1	0	1	2
29	J45.9	asma	30-300	60-300	1	1	1	3
29	K57.9	DIVERTICULOSIS	30-300	30-300	1	1	1	3

29	K76.0	esteatosi hepàtica	30-300	30-300	1	1	0	2
29	N39.4	Incontinència urinària i fecal	30-300	30-300	0	1	0	1
29	T30.0	cremada	0-15	0-15	0	0	0	0
28	E78.0	HIpercokesterolèmia	30-300	30-300	0	1	1	2
28	H40.9	galucoma	30-300	30-300	0	0	1	1
28	M41.9	cifoescoliosis	30-300	30-300	0	1	0	1
28	N42.9	PROSTATITIS CRONICA	30-300	30-300	0	1	0	1
27	A48.3	sindrome toxico	5-120	5-120	0	0	1	1
27	D12.6	POLIPO PEDICULADO	30-300	30-300	0	1	0	1
27	I21.1	IAM inferior	5-300	1-60	1	1	1	3
27	I51.9	Cardiopatia crònica	30-300	30-300	1	1	1	3
27	I70.9	ATEROMATOSIS CAROTIDEA BILATERAL CON ESTENOSIS	30-300	30-300	0	1	1	2
27	K20	esofagitis En tercio medio esofagico	30-300	30-300	0	0	1	1
27	M10.9	crisi gotosa	5-30	5-30	0	1	1	2
27	R33	RAO	5-30	5-30	0	0	1	1
27	R47.1	disartria	5-60	5-30	0	0	1	1
26	E10	diabetis i	30-300	30-300	1	1	1	3
26	I87.2	insuficiencia venosa	10-120	10-120	0	0	0	0
26	M13.9	Artritis Aguda	5-30	5-30	0	0	1	1
25	C78.7	metàstatsis al fetge	30-100	30-300	1	1	1	3
25	H92.0	Otalgia derecha	5-30	5-30	0	0	0	0
25	I35.1	DOBLE LESION AORTICA MODERADA	30-300	30-300	0	1	0	1
25	I83.9	lligadura de varius	30-300	30-300	0	0	0	0
25	J98.0	Broncoespasme generalitzat	5-60	1-5	0	0	1	1
25	M51.9	discopatia L5-S1	30-300	30-300	0	1	0	1
25	M54.4	LUMBOCIATALGIA DRETA	5-120	5-120	0	0	0	0
24	C80	Adenocarcinoma pulmopnar	30-300	30-300	0	1	1	2
24	F52.2	disfunció erèctil	5-120	5-120	0	1	0	1
24	Z92.9	AMC	30-300	30-300	0	0	0	0
23	I21	IAM no Q Killip I	30-300	1-60	1	1	1	3
23	I45.0	BBD	30-300	30-300	0	0	0	0
23	K85.9	pancreatitis	30-300	1-60	1	0	1	2
23	T14.3	entorsis de turmell E	5-15	5-15	0	0	0	0
22	B35.1	onicomicosis	30-300	30-300	0	0	0	0

22	C78.0	METASTASIS PULMONARS	30-300	30-180	1	1	1	3
22	I25.3	aurícula izquierda aneurismática	30-300	30-180	0	0	1	1
22	I50	reagudización de IC	5-60	5-60	0	0	1	1
22	J96.9	IR	15-120	15-120	0	0	1	1
22	R59.9	Adenopatia	5-15	5-15	0	0	0	0
22	Z13.60	FRCV	30-300	30-300	0	0	0	0
21	B00.9	herpes	5-30	5-30	0	0	0	0
21	B02.9	HERPES ZOSTER	5-30	5-30	1	1	0	2
21	M54.1	Raquialgias crónicas	30-300	30-300	0	1	0	1
20	A04.4	Ecoli	5-15	5-15	0	0	0	0
20	E05.9	HIPERTIROIDISMO	30-300	30-180	1	1	1	3
20	E28.3	menopausia	30-300	30-300	0	0	0	0
20	I25	enfermedad de un vaso	30-300	30-300	0	0	1	1
20	I35.0	estenosis	30-300	30-300	0	0	1	1
20	J43.9	EMFISEMA	30-300	30-300	1	1	1	3
20	K90.4	intolerancia	30-300	30-300	0	0	0	0
20	L82	M óssies	30-300	30-180	0	0	1	1
20	W57.9	picada de insecto	5-10	5-10	0	0	0	0
20	Z93.2	ileostomía	30-300	30-300	0	1	1	2
19	C79.5	Metástasis óseas	30-300	30-180	1	1	1	3
19	N08.3	nefropatia diabètica	30-300	30-300	1	1	1	3
19	Z88	alergias medicamentosas conocidas	30-300	30-300	0	0	0	0
18	B37.9	candidiasisi	5-60	5-60	0	0	0	0
18	C79.3	METASTASIS CEREBRALS	30-300	30-180	1	1	1	3
18	D17.9	lipoma	10-300	10-300	0	0	0	0
18	F51.0	INSOMNI	10-100	10-100	0	0	1	1
18	G93.1	Leucoencefalopatía hipóxica subcortical crónica	30-300	30-300	0	1	0	1
18	I38	valvulopatias	30-300	30-300	0	1	1	2
18	J42	Bronquitis crònica	30-300	30-300	1	1	1	3
18	M20.1	HALLUX	30-300	30-300	0	0	0	0
17	D69.6	Trombocitosis reactiva	30-300	30-300	0	0	1	1
17	E87.6	hipok inferobasal con isquemia asociada	30-300	30-300	0	1	1	2
17	G35	EM lleu	30-300	30-300	0	0	0	0
17	H11.3	Hiposfagma OI	5-120	5-120	0	0	0	0
17	I61.9	AVC hemoragic	5-120	1-15	1	1	1	3
17	J96.0	insuficiencia respiratoria aguda	5-30	1-5	1	0	1	2

17	M47.8	Lumbartrosis	30-300	30-300	0	1	1	2
17	M47.9	espondiloartrosis lumbar	30-300	30-300	0	1	1	2
16	E79.0	Hiperuricemia	30-300	30-300	0	0	0	0
16	H60.9	otitis	5-20	5-20	0	0	0	0
16	H81.3	sindrome vestibular periferic esquerre	10-300	10-300	0	1	1	2
16	I27.2	HTp severa	30-300	30-300	0	1	1	2
16	M23.3	meniscopatia	10-300	10-300	0	1	0	1
16	M79.7	fibromiàlgia	30-300	90-180	1	1	1	3
16	N23	tipus còlic	10-60	10-60	0	0	0	0
16	R80	microalbuminurua	10-120	10-120	0	0	0	0
16	Z54	convalescència	30-300	30-300	0	0	0	0
16	Z63.4	dol	10-300	10-300	0	0	0	0
15	B37.0	candiasis oral	5-20	5-20	0	0	0	0
15	D18.0	hemangioma	30-300	30-300	0	0	0	0
15	H65.9	otitis serosa	5-20	5-20	0	0	0	0
15	I25.8	Microinfarto isquémico subagudo	10-120	10-120	0	1	1	2
15	I26.9	MALALTIA TROMBOEMBÓLICA PARANEOPLASICA	30-300	30-300	0	1	1	2
15	I49.9	arritmia	10-120	10-120	0	0	1	1
15	I84.2	Hemorroides internas	5-30	5-30	0	0	0	0
15	J31	Rinitis senil	30-300	30-300	0	0	0	0
15	J90	vessament pleural	5-120	15-30	1	0	1	2
15	K76.9	HEPATOPATÍA CRONICA	30-300	30-300	1	0	1	2
15	L21.9	DERMATITIS SEBORREICA	30-120	30-120	0	0	0	0
15	L57.0	queratosis actinica	5-30	5-30	0	0	0	0
15	M47	espondilosis dorsal	30-300	30-300	0	0	0	0
15	M61.4	calcificaio	30-300	30-300	0	0	0	0
15	M79.6	Omàlgia D	5-120	5-120	0	0	0	0
15	R00.0	taquicardia sinusal	10-120	10-120	0	0	0	0
15	Z11	tolerancia de dogmatil	30-300	30-300	0	0	0	0
14	E04.2	GOLL MULTINODULAR EUTIROIDEO	30-300	30-300	0	1	0	1
14	I05.0	estenosi	5-120	5-120	0	0	0	0
14	I36.1	IT ligera	5-30	5-30	0	0	0	0
14	K52.9	DIARREA	5-15	5-15	0	0	0	0
14	M71.9	bursitis olecranon derecha	5-120	5-120	0	0	0	0
14	R40.0	Somnolencia	5-30	5-30	0	0	0	0
14	Y83.5	amputación	5-120	5-120	0	0	1	1

14	Z90.4	COLECISTECTOMIA	30-300	30-300	0	0	0	0
13	D64.8	anemia macrocitica hipercromica	30-300	30-300	0	1	1	2
13	E78.1	hipertrigliceridemia	30-300	30-300	0	1	0	1
13	E87.5	hiperpotassèmia	15-120	15-120	0	1	0	1
13	G56.0	tunel carpià bilateral	30-300	30-300	0	1	0	1
13	G62.9	polineuropatia sensitiva - motora axonal	30-300	30-300	0	1	1	2
13	H35.3	degeneracio macular amb drusses	30-300	30-180	1	0	1	2
13	H53.2	diplopia	5-15	1-15	0	0	1	1
13	J01.9	sinusitis frontal	5-30	5-30	1	0	0	1
13	K60.2	fissura anal	5-30	5-30	0	1	0	1
13	L97	ulcera	30-120	5-60	1	0	1	2
13	M48.2	cervicoartrosis severa	30-300	30-300	0	1	0	1
13	M79.2	alodinia	30-120	30-120	0	0	1	1
13	R29.1	meningismo	0-20	0-20	0	0	1	1
13	Z22.5	VHC	30-300	60-180	1	0	1	2
13	Z73.6	Limitació funcional d'ambdues extremitats	30-300	30-300	0	0	0	0
12	A41.9	sepsis abdominal	5-30	5-30	0	0	1	1
12	B96.8	h. Pylori	15-120	15-120	1	1	1	3
12	C67.9	neo maligna de pared lateral vejiga urinaria	30-300	30-300	0	1	1	2
12	E14.5	PEU DIABETIC	30-300	30-300	0	1	1	2
12	H81.9	sdme vertiginoso	5-30	5-30	0	0	1	1
12	J06.9	IRVB	5-30	5-30	1	0	1	2
12	M32.9	Lupus	30-300	30-300	1	1	1	3
12	M51.2	HERNIA DISCAL	30-300	30-300	1	1	0	2
12	M70.6	Troncanteritis izq	5-30	5-30	0	1	0	1
12	R06.8	encefalopatia hipercàpnica	10-120	10-120	0	1	1	2
11	B07	berruga a cuixa esquerra	15-120	15-120	0	0	0	0
11	B97.7	cervicopatia	30-300	30-300	0	0	0	0
11	F01.9	demencia vascular	30-300	30-300	0	1	1	2
11	G47.3	apnea del son	30-300	30-300	0	0	1	1
11	I24.9	isquemia aguda	5-15	5-15	0	0	1	1
11	I25.5	Miocardiopatia isquémica	30-300	30-180	1	1	1	3
11	I77.5	Necrosis inferior	30-300	30-300	0	1	1	2
11	J47	Bronquiectasi	30-300	30-300	0	1	1	2
11	J98.4	ALTRES TRASTORNS DEL PULMÓ	5-300	5-300	0	0	0	0
11	K12.0	aftosis bocal	5-30	5-30	0	0	0	0
11	K55.1	colitis isquèmica	30-300	60-300	1	1	1	3

11	M40.5	hiperlordosis lumbar	30-300	30-300	0	0	0	0
11	M77.3	Espolon calcaneo bilateral	30-120	30-120	0	0	0	0
11	R20.1	hipoestesia a nivell mal.leolar interna del peu esquerra	30-120	30-120	0	0	1	1
11	R47	AFÀSIA	30-300	1-5	0	0	1	1
11	R53	Astènia	30-300	30-300	0	0	0	0
10	C16.9	CÀNCER GÀSTRIC SUBCARDIAL	30-300	30-180	0	1	1	2
10	C18.7	ca sigma in situ	30-150	30-180	0	1	1	2
10	C73	hiperplasia paratiroides	30-300	30-300	0	1	1	2
10	F10.2	enolisme crònic	30-300	30-300	1	1	1	3
10	G53.0	neuralgia postherpetica	30-120	30-120	1	1	1	3
10	I27.0	HAP severa	30-300	30-300	0	1	1	2
10	I63	infarto subagudo en AICA derecha	5-15	5-15	1	1	1	3
10	I67.8	isquemia	5-120	5-120	0	0	1	1
10	I82.9	trombosis en la zona de las venas profundas de la pantorrilla	5-120	5-120	0	0	1	1
10	K40.9	hèrnia inguinal	30-300	30-300	0	1	1	2
10	L40.9	PSORIASIS	30-300	30-300	1	1	1	3
10	L60.0	Ungla del peu esquerra encarnata	10-120	10-120	0	0	0	0
10	N18.0	I. Renal terminal	30-300	30-300	0	0	1	1
10	S29	sobreinfección	5-30	5-30	0	0	1	1
10	Z86	anteced similares	5-300	5-300	0	0	0	0
9	F43.2	tr. Adaptativo	30-300	30-300	1	0	0	1
9	G63.2	polineuropatia diabètica	30-300	30-300	1	1	1	3
9	H81.1	vertigo poicional benigno	5-30	5-30	0	1	1	2
9	J03.9	AMIGDALITIS AGUDA	5-10	5-10	0	0	0	0
9	J32.9	sinusitis cronica	30-300	30-300	1	1	0	2
9	K05.4	periodontitis	5-120	5-120	0	0	0	0
9	K56.6	suboclusion intestinal	5-30	5-30	0	0	1	1
9	L02.9	absceso en region mamaria	5-30	5-30	0	0	0	0
9	M99.3	estenosi del canal raquidi	30-300	30-300	0	1	0	1
9	R04.0	epistaxis	5-15	5-15	0	0	0	0
9	R06.0	DPN			0	0	1	1
9	R06.7	rinorrea	5-15	5-15	0	0	0	0
9	R13	Disfàgia	30-120	15-180	0	0	1	1
9	R47.0	afasia leve	5-120	5-120	0	0	1	1
9	R50.9	Síndrome Febril Agudo	5-30	5-30	0	0	0	0
9	Z04.3	accident	5-30	5-30	0	0	0	0

8	B34.9	viriasis	5-15	5-15	0	0	0	0
8	B86	Escabiosis	5-60	5-60	0	0	0	0
8	D23.9	tumor benigne cutani	30-120	30-120	0	0	0	0
8	E86	DESHIDRATACIO	5-30	5-30	0	0	0	0
8	F01.1	Cadasil			0	0	0	0
8	G45.8	leisons isquemiques agudes	5-30	5-30	0	0	0	0
8	I11	cardiopatia HTA	30-300	30-300	0	0	0	0
8	I35.2	EAo moderada	30-300	30-300	0	0	0	0
8	I42.9	Miocardiortpatia	30-300	30-300	0	0	1	1
8	I44.0	BAV 1er grado	30-150	30-150	0	0	0	0
8	I50.1	ICI NHYA II	30-300	30-300	0	0	1	1
8	J98.1	colapse pulmonar	5-30	1-5	0	0	0	0
8	L63.9	alopecia generalizada	30-300	30-300	0	0	0	0
8	L85.9	hiperqueratosi plantar	30-300	30-300	0	0	0	0
8	M13.2	LUXACION CABEZA FEMUR EN ACCIDENTE TRAFICO	5-30	5-30	0	0	0	0
8	M15.9	POLIARTRALGIAS	30-300	30-300	0	0	0	0
8	M25.9	ARTROPATIA DEG GENERALIZADA	30-300	30-300	0	0	0	0
8	M31.9	Vasculopatia periferica	10-120	10-120	0	0	1	1
8	N28.9	nefropatia	30-300	30-300	0	0	1	1
8	R12	Pirosi	30-300	30-300	0	0	0	0
8	R60.9	augment edemes	30-120	30-120	0	0	0	0
8	R62.8	Sd contitucional	30-120	30-120	0	0	0	0
8	R94.2	Hiperreactivitat bronquial	5-120	5-120	0	0	0	0
8	Z96.1	pseudofaquia OI	30-300	30-300	0	0	0	0
7	C50.9	Ca.Mama	30-300	90-300	0	0	1	1
7	C92.1	Leucemia Mieloide crònica	30-150	30-150	0	0	1	1
7	E10.9	Insulinodepenent	30-150	30-150	0	0	0	0
7	G40.1	epilepsia parcial	30-300	30-300	0	0	0	0
7	G91.9	HIDROCEFALIA	30-300	30-300	0	0	0	0
7	J11	Gripe	5-30	5-30	0	0	0	0
7	J42	BRONQ			0	0	0	0
7	K80.1	colecistitis	30-300	2-30	1	0	1	2
7	K80.2	patología biliar	30-300	30-300	1	0	0	1
7	L20.9	ANGOR	5-120	15-90	1	0	1	2
7	M13.8	artritis gotosa	30-300	30-300	1	0	0	1
7	M25.7	osteofitos	30-300	30-300	0	0	0	0
7	M35.3	POLIMIALGIA	30-120	30-120	0	0	0	0
7	M65.9	Tenosinovitis de porción larga del biceps	5-30	5-30	0	0	0	0
7	M72.0	Dupuytren	30-300	30-300	0	0	0	0

7	N12	PIELONEFRITIS IZQUIERDA	30-300	30-300	0	0	0	0
7	Q61.3	poliquistois renal	30-300	30-300	0	0	0	0
7	R04.2	hemoptisi	5-60	5-60	0	0	0	0
7	R14	flatulència	30-120	30-120	0	0	0	0
7	R23.0	cianosi	5-30	5-30	0	0	0	0
7	R74.0	hipertransaminassèmia	30-300	30-300	1	0	0	1
7	T00.9	contusiones	5-120	5-120	0	0	0	0
6	A04	ENTEROCOLITIS	5-30	5-30	0	0	0	0
6	B18.2	HEPATITIS C	30-150	30-180	1	0	1	2
6	C41.9	M1		30-180	0	0	1	1
6	C96.9	LINFOMA	30-300	30-180	0	0	1	1
6	D36.1	NEURINOMA	30-300	30-300	0	0	1	1
6	D57.1	drepanocitosis	30-300	30-300	0	0	0	0
6	E14.1	descompensacion hiperglicemica	5-60	5-60	0	0	0	0
6	E66.2	hipoventilacio alveolar	10-120	10-120	0	0	0	0
6	E78.2	Dislipemia mixta	30-300	30-300	1	0	0	1
6	E83.5	hipercalcemia secundarioa a farmacos	15-120	15-120	0	0	0	0
6	F10.1	enolismo	30-300	90-300	1	0	0	1
6	G40.9	epilepsia cortical focal motora	30-300	30-300	0	0	1	1
6	G72.9	miopatia	30-120	90-300	1	0	0	1
6	H00.0	orzuelo en parpado inferior ojo derecho	5-30	5-30	0	0	0	0
6	H02.4	Ptosis palpebral	5-150	5-150	1	0	0	1
6	H53.4	HEMIANOPSIA OD	5-150	5-150	0	0	1	1
6	H54.7	VISIÓ SUBNORMAL D'UN ULL	5-30	5-30	0	0	0	0
6	H66.9	otitis media	5-30	5-30	1	0	0	1
6	I21.0	IAM anterior	5-60	5-60	1	0	1	2
6	I44.2	BAV completo	30-150	30-150	1	0	0	1
6	I50.91	disfunción diastólica	30-150	30-150	0	0	0	0
6	I60.9	HSA	5-120	5-120	0	0	1	1
6	I74.9	Primera Diagonal con lesión de aspecto trombótico	5-60	5-60	0	0	0	0
6	I89.0	Linfedema	5-150	5-150	0	0	0	0
6	I95.1	HIPOTENSIÓ ORTOSTÀTICA	5-120	5-120	0	0	0	0
6	J30.4	rinitis alergica	30-150	30-150	1	0	0	1
6	J81	EPA			0	0	0	0

6	K04.9	flegmó dentari	5-30	5-30	0	0	0	0
6	K14.0	glositis	5-30	5-30	0	0	0	0
6	K28.9	ulcus duodenal	30-150	30-180	1	0	1	2
6	K29.7	gastritis	30-150	30-150	0	0	0	0
6	K43.9	Eventración abdominal	30-150	30-150	0	0	0	0
6	K62.8	parálisis del recto lateral derecho	5-120	5-120	0	0	0	0
6	K76.8	quistes hepaticos	30-300	30-300	0	0	0	0
6	K85	Pancreatitis aguda litiàsica	5-30	2-30	1	0	0	1
6	M72.2	FASCITIS PLANTAR BILATERAL	30-300	30-300	0	0	0	0
6	M75.1	Sd del manegot dels rotadors	30-300	30-300	0	0	0	0
6	M77.1	EPICONDILITIS IZDA	5-30	5-30	0	0	0	0
6	N48.1	balanitis per antibioterapia	5-30	5-30	0	0	0	0
6	N81.4	prol. Ueri	5-60	5-60	0	0	0	0
6	Q62.3	ectasias	5-60	5-60	0	0	0	0
6	R11	nàusees	5-30	5-30	0	0	0	0
6	R31	Hematuria	5-60	30-180	1	0	0	1
6	R45.8	trauma	5-30	5-30	0	0	0	0
6	R58	Gingivorragias	5-30	5-30	0	0	0	0
6	S19	infecció	5-30	5-30	0	0	0	0

Table 28: Recommended time gap and importance per diagnosis

B Composition Pack 1

The following table shows, the composition of the first pack of rules. *TC* is the average time between antecedents and consequent and *TA* is the average time between antecedents. *Sup* represents the *Support*, *Con* the *Confidence*, *Lif* the *Lift*, and *X* the *rulesRespectConsequent* indicator. The results can be observed in the next page.

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
1	[’AGENTES ANTITROMBÓTICOS’]	[’Infarto agudo del miocardio’]	15,50	0,00	0,0031	0,0263	1,8413	0,2177
2	[’Enfermedad isquémica crónica del corazón’]	[’Infarto agudo del miocardio’]	14,50	0,00	0,0022	0,0731	5,113	0,1532
3	[’OTROS ANALGÉSICOS Y ANTIPIRÉTICOS’]	[’Infarto agudo del miocardio’]	15,60	0,00	0,0025	0,0185	1,2979	0,1774
4	[’ANTISÉPTICOS Y DESINFECTANTES’, ’DOLOR GENERALIZADO/MÚLTIPLE’, ’CAMBIOS EN EL COLOR DE LA PIEL’]	[’Traumatismo de regiones no especificadas del cuerpo’]	13,70	5,50	0,0022	0,2346	6,3997	0,0597
5	[’INFECCIÓN DERMATOLÓGICA POSTRAUMÁTICA’, ’ANTISÉPTICOS Y DESINFECTANTES’]	[’Traumatismo de regiones no especificadas del cuerpo’]	12,90	4,40	0,0025	0,2037	5,5577	0,0692
6	[’SIGNOS/SÍNTOMAS DE LA TEXTURA CUTÁNEA’, ’ANTISÉPTICOS Y DESINFECTANTES’]	[’Traumatismo de regiones no especificadas del cuerpo’]	13,20	4,10	0,0023	0,1282	3,4978	0,0629
7	[’DOLOR GENERALIZADO/MÚLTIPLE’, ’CAMBIOS EN EL COLOR DE LA PIEL’]	[’Traumatismo de regiones no especificadas del cuerpo’]	15,90	2,80	0,0028	0,1348	3,6786	0,0755
8	[’ANTISÉPTICOS Y DESINFECTANTES’, ’ANTIBACTERIANOS PARA USO SISTEMICO’]	[’Traumatismo de regiones no especificadas del cuerpo’]	12,70	5,90	0,0021	0,125	3,4104	0,0566
9	[’OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO’, ’OTROS SIGNOS/SÍNTOMAS NEUROLÓGICOS’]	[’Traumatismo de regiones no especificadas del cuerpo’]	18,20	2,80	0,0022	0,0512	1,3972	0,0597

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
10	['TOS', 'FIEBRE']	['Neumonía, organismo no especificado']	15,80	1,30	0,0021	0,0434	6,7198	0,3214
11	['FIEBRE', 'FATIGA RESPIRATORIA/DISNEA', 'PALPITACIONES/PERCEPCIÓN DE LOS LATIDOS CARDIACOS']	['Insuficiencia cardíaca']	16,50	5,50	0,0031	0,1971	13,9014	0,2195
12	['Otras enfermedades pulmonares obstructivas crónicas', 'FATIGA RESPIRATORIA/DISNEA']	['Insuficiencia cardíaca']	16,80	6,20	0,0031	0,1837	12,9557	0,2195
13	['TOS', 'PALPITACIONES/PERCEPCIÓN DE LOS LATIDOS CARDIACOS']	['Insuficiencia cardíaca']	15,50	6,00	0,0031	0,18	12,6966	0,2195
14	['DIURÉTICOS DE TECHO ALTO']	['Insuficiencia cardíaca']	12,50	0,00	0,0038	0,0766	5,4007	0,2683
15	['RESPIRACIÓN JADEANTE/SIBILANTE', 'FIEBRE']	['Insuficiencia cardíaca']	16,80	5,40	0,0032	0,1647	11,6178	0,2276
16	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'OTRAS IRREGULARIDADES DEL RITMO CARDÍACO', 'FATIGA RESPIRATORIA/DISNEA']	['Insuficiencia cardíaca']	15,90	5,40	0,0031	0,1617	11,4041	0,2195
17	['OTROS PROBLEMAS DE LA RESPIRACIÓN', 'Insuficiencia cardíaca', 'PALPITACIONES/PERCEPCIÓN DE LOS LATIDOS CARDIACOS']	['Otras enfermedades pulmonares obstructivas crónicas']	16,10	5,50	0,0031	0,3553	19,508	0,1709

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
18	['OTROS PROBLEMAS DE LA RESPIRACIÓN', 'CORTICOSTEROIDES PARA USO SISTÉMICO, MONOTERAPIA']	['Otras enfermedades pulmonares obstructivas crónicas']	15,40	4,60	0,0035	0,2857	15,689	0,1899
19	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'RESPIRACIÓN JADEANTE/SIBILANTE', 'FATIGA RESPIRATORIA/DISNEA', 'OTROS PROBLEMAS DE LA RESPIRACIÓN']	['Otras enfermedades pulmonares obstructivas crónicas']	16,50	6,70	0,0033	0,2566	14,0923	0,1835
20	['Atrovent', 'FATIGA RESPIRATORIA/DISNEA']	['Otras enfermedades pulmonares obstructivas crónicas']	15,20	4,00	0,0031	0,2547	13,9869	0,1709
21	['RESPIRACIÓN JADEANTE/SIBILANTE', 'Salbutamol']	['Otras enfermedades pulmonares obstructivas crónicas']	15,30	3,20	0,0031	0,2195	12,0537	0,1709
22	['Insuficiencia cardíaca', 'FATIGA RESPIRATORIA/DISNEA']	['Otras enfermedades pulmonares obstructivas crónicas']	16,50	6,40	0,0033	0,2028	11,1359	0,1835
23	['RESPIRACIÓN JADEANTE/SIBILANTE', 'FATIGA RESPIRATORIA/DISNEA']	['Otras enfermedades pulmonares obstructivas crónicas']	15,80	4,70	0,0043	0,1937	10,6373	0,2342
24	['RESPIRACIÓN JADEANTE/SIBILANTE', 'FIEBRE']	['Otras enfermedades pulmonares obstructivas crónicas']	18,10	5,00	0,0037	0,1882	10,3363	0,2025

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
25	[’TOS’, ’FIEBRE’, ’FATIGA RESPIRATORIA/DISNEA’]	[’Otras enfermedades pulmonares obstructivas crónicas’]	16,00	4,90	0,0036	0,1512	8,3037	0,1962

Table 29: First pack results

C Composition Pack 2

The following table shows, the composition of the second pack of rules. *TC* is the average time between antecedents and consequent and *TA* is the average time between antecedents. *Sup* represents the *Support*, *Con* the *Confidence*, *Lif* the *Lift*, and *X* the *rulesRespectConsequent* indicator. The results can be observed in the next page.

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
1	['OTROS SIGNOS/SÍNTOMAS NEUROLÓGICOS']	['Infarto cerebral']	63,90	0,00	0,0020	0,0110	1,5033	0,2747
2	['PRODUCTOS ANTIINFLAMATORIOS Y ANTIRREUMATICOS NO ESTEROIDEOS']	['Otros trastornos de los tejidos blandos']	65,80	0,00	0,0024	0,0196	1,7687	0,2174
3	['OTROS ANALGÉSICOS Y ANTIPIRÉTICOS', 'DOLOR GENERALIZADO/MÚLTIPLE']	['Otros trastornos de los tejidos blandos']	69,50	22,80	0,0021	0,0139	1,2512	0,1884
4	['Neoplasia maligna']	['Tumores [neoplasias] malignos de sitios mal definidos, secundarios y de sitios no especificados']	56,70	0,00	0,0035	0,0304	3,9796	0,4526
5	['OTROS ANALGÉSICOS Y ANTIPIRÉTICOS']	['Tumores [neoplasias] malignos de sitios mal definidos, secundarios y de sitios no especificados']	57,90	0,00	0,0031	0,0119	1,5545	0,4000
6	['DOLOR GENERALIZADO/MÚLTIPLE', 'EEII']	['Flebitis y tromboflebitis']	53,30	20,10	0,0020	0,0159	2,4142	0,3049
7	['OTROS SIGNOS/SÍNTOMAS DEL APARATO LOCOMOTOR']	['Flebitis y tromboflebitis']	60,30	0,00	0,0022	0,0157	2,3869	0,3293
8	['DOLOR GENERALIZADO/MÚLTIPLE', 'OTROS SIGNOS/SÍNTOMAS CARDIOVASCULARES']	['Flebitis y tromboflebitis']	54,20	21,90	0,0020	0,0153	2,3139	0,3049
9	['PREPARADOS CON HIERRO']	['Otras anemias']	67,20	0,00	0,0023	0,0903	6,8054	0,1697
10	['ASTENIA/ CANSANCIO/ DEBILIDAD GENERAL']	['Otras anemias']	41,00	0,00	0,0020	0,0248	1,8705	0,1515

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
11	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'OTROS SIGNOS/SÍNTOMAS CARDIOVASCULARES']	['Otras anemias']	55,00	11,80	0,0022	0,0168	1,2628	0,1636
12	['Hipertensión esencial (primaria)']	['Otras anemias']	71,90	0,00	0,0023	0,0156	1,1754	0,1758
13	['Neumonía, organismo no especificado']	['Neoplasia maligna']	58,90	0,00	0,0023	0,1062	3,8390	0,0843
14	['DIARREA', 'OTROS ANALGÉSICOS Y ANTIPIRÉTICOS', 'abdominal']	['Neoplasia maligna']	58,20	11,50	0,0020	0,0731	2,6418	0,0727
15	['CAMBIO EN LAS HECES/EN EL RITMO INTESTINAL']	['Neoplasia maligna']	65,30	0,00	0,0026	0,0603	2,1779	0,0930
16	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'OTROS ANALGÉSICOS Y ANTIPIRÉTICOS', 'TOS', 'DOLOR GENERALIZADO/MÚLTIPLE']	['Neoplasia maligna']	59,00	29,10	0,0020	0,0571	2,0628	0,0727
17	['OTROS ANALGÉSICOS Y ANTIPIRÉTICOS', 'Pazital']	['Neoplasia maligna']	57,80	15,40	0,0033	0,0566	2,0466	0,1192
18	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'OTROS SIGNOS/SÍNTOMAS NASALES']	['Neoplasia maligna']	60,70	22,20	0,0020	0,0543	1,9641	0,0727
19	['OTROS ANALGÉSICOS Y ANTIPIRÉTICOS', 'DOLOR GENERALIZADO/MÚLTIPLE', 'abdominal']	['Neoplasia maligna']	59,70	21,00	0,0032	0,0498	1,8002	0,1163

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
20	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'TOS', 'DOLOR GENERALIZADO/MÚLTIPLE']	['Neoplasia maligna']	56,20	29,40	0,0025	0,0476	1,7209	0,0901
21	['ASTENIA/ CANSANCIO/ DEBILIDAD GENERAL']	['Neoplasia maligna']	62,90	0,00	0,0038	0,0467	1,6868	0,1366
22	['OTROS SIGNOS/SÍNTOMAS DEL APARATO RESPIRATORIO', 'TOS', 'FIEBRE']	['Neoplasia maligna']	59,10	19,10	0,0032	0,0460	1,6616	0,1163
23	['Trastornos mentales y del comportamiento debidos al uso de tabaco']	['Neoplasia maligna']	59,20	0,00	0,0026	0,0419	1,5157	0,0930
24	['EXPECTORANTES, EXCLUIDOS COMBINACIONES CON SUPRESORES DE LA TOS']	['Neoplasia maligna']	59,30	0,00	0,0021	0,0357	1,2907	0,0756
25	['SIGNOS/SÍNTOMAS DE LA TEXTURA CUTÁNEA', 'ANTISÉPTICOS Y DESINFECTANTES', 'ANTIBACTERIANOS BETALACTÁMICOS, PENICILINAS']	['Atencio domiciliaria']	53,30	21,80	0,0029	0,4865	27,6164	0,1644
26	['ENZIMAS', 'Traumatismo de regiones no especificadas del cuerpo']	['Atencio domiciliaria']	54,20	9,40	0,0029	0,2791	15,8420	0,1644
27	['SIGNOS/SÍNTOMAS DE LA TEXTURA CUTÁNEA', 'DOLOR GENERALIZADO/MÚLTIPLE', 'FIEBRE']	['Atencio domiciliaria']	53,50	23,50	0,0029	0,2156	12,2372	0,1644

ID	Antecedent	Consequent	TC	TA	Sup	Con	Lif	X
28	[’FIEBRE’, ’ANTIBACTERIANOS PARA USO SISTEMICO’, ’Traumatismo de regiones no especificadas del cuerpo’]	[’Atencio domiciliaria’]	56,50	20,10	0,0029	0,2156	12,2372	0,1644
29	[’ANTISÉPTICOS Y DESINFECTANTES’, ’ANTIBACTERIANOS BETALACTÁMICOS, PENICILINAS’]	[’Atencio domiciliaria’]	58,40	22,60	0,0032	0,2041	11,5851	0,1826
30	[’ANTISÉPTICOS Y DESINFECTANTES’, ’DOLOR GENERALIZADO/MÚLTIPLE’, ’ANTIBACTERIANOS PARA USO SISTEMICO’]	[’Atencio domiciliaria’]	56,60	24,50	0,0032	0,1923	10,9168	0,1826
31	[’ANTISÉPTICOS Y DESINFECTANTES’, ’DOLOR GENERALIZADO/MÚLTIPLE’, ’CAMBIOS EN EL COLOR DE LA PIEL’]	[’Atencio domiciliaria’]	54,60	20,60	0,0031	0,1674	9,5018	0,1781
32	[’Traumatismo de regiones no especificadas del cuerpo’, ’ANTIBACTERIANOS BETALACTÁMICOS, PENICILINAS’]	[’Atencio domiciliaria’]	52,70	17,60	0,0029	0,1412	8,0142	0,1644
33	[’dits ma D’, ’DOLOR GENERALIZADO/MÚLTIPLE’]	[’Atencio domiciliaria’]	67,30	17,30	0,0029	0,0870	4,9363	0,1644
34	[’ERITEMA/RASH LOCALIZADO’]	[’Atencio domiciliaria’]	49,90	0,00	0,0032	0,0496	2,8137	0,1826

Table 30: Second pack results

D Evaluations Pack 1

The following table shows, the evaluation of each doctor for the first pack of rules. *C* represents the *Correctness*, *R* the *Relevance* of a rule. Regarding the letters inside the brackets, *L* refers to *Luis*, *D* to *David* and *V* to *Victor*. The ID of the rules correspond to the ones presented in Table 29. For visualization purposes, we also abbreviated the following labels:

- Correctness:
  - I: Totally incorrect.
  - C: Totally correct.
  - P-T: Partially correct (for the temporal aspect).
  - P-C: Partially correct (for the clinical aspect).
  - P-B: Partially correct (for both aspects).
- Relevancy:
  - N: Not relevant.
  - R-K: Relevant and known.
  - R-U: Relevant and unknown.

The results can be observed in the next page.

ID	C(V)	C(D)	C(L)	R(V)	R(D)	R(L)	Justification (V)	Justification (D)	Justification (L)
1	C	P-C	P-C	R-K	R-K	N	És un dels tractaments de l'IAM	Associació evident, però no única associació possible	Se li va pautar un antitrombòtic per un altre patologia
2	C	C	C	R-K	N	R-K	Estan molt relaciones són part de la mateixa malatia.	mateix codi	Hi ha molta evidencia científica que estableix una relació rellevant
3	P-C	I	P-C	R-K	N	N	Si s'inclou l'A.A.S com a antipirètic és cert, si són altres sería finsi tot rellevant i sorprenent	probablement el codi de fàrmac fa referència a l'AAS, catalogat com analgèsic i antipirètic	L'infart pot anar precedit de dolor al pit que porti al pacient a pendre analgésics
4	C	P-C	I	R-K	R-K	N	Són les cures habituals en un traumatisme i els canvis posteriors en la cicatriu del traumatisme.	Sí relació	No veig cap relació
5	C	P-C	P-C	R-K	R-K	N	Complicació habitual i són les cures habituals en un traumatisme	Sí relació	Tenir un trauma predisposa tenir-ne un altre
6	C	P-C	I	N	R-K	N	S'entén que és la crosta d'una ferida	Sí relació	No trobo cap relació
7	C	P-C	P-C	R-K	R-K	R-K	És normal tenir dolor en un traumatisme i els canvis posteriors en la cicatriu del traumatisme.	Relació menys evident	Dolor + hematoma – >Predisposa a un altre trauma
8	C	P-C	I	R-K	R-K	N	Complicació habitual i tractament habitual, potencialment perillòs.	Sí relació	No trobo cap relació

ID	C(V)	C(D)	C(L)	R(V)	R(D)	R(L)	Justification (V)	Justification (D)	Justification (L)
9	C	I	P-C	R-K	N	N	Tant la falt d'aires, tos, i altres símptomes poden fer un sícop i fer un traumatisme	No hauria de tenir cap relació	Determinats signes neurologics favoreixen caigudes i traumas
10	C	C	C	R-K	R-K	R-K	Diagnòstic típic	Relació clara i coneguda	La majoria de penumonias ven precedides de tos i/o febre
11	P-C	C	P-C	R-K	R-K	R-K	Es plausible que en el context de sobreinfecció respiratòria faci una insuf cardíaca, però no és el diagnòstic esperat per la presència de febre	Relació possible en alguns casos	Son signes que acaban en un diagnóstic de IC
12	C	C	P-C	R-K	R-K	R-K	Es pot entendre que el pacient presenta primer MPOC i s'afegeixi més díspnea i siguui d'orígen cardíac	Relació possible en alguns casos	Son signes que acaban en un diagnóstic de IC
13	C	C	C	R-K	R-K	R-K	Símptomes típics	Relació possible en alguns casos	Son signes que acaban en un diagnóstic de IC
14	C	C	I	R-K	R-K	R-K	És un tractament típic	Relació clara i coneguda	Retenció hídrica com a consequencia de fases incials no diagnosticades de IC
15	P-C	C	P-C	R-K	R-K	R-K	Es plausible que en el context de sobreinfecció respiratòria faci una insuf cardíaca, però no és el diagnòstic esperat	Relació possible	Son signes que acaban en un diagnóstic de IC

ID	C(V)	C(D)	C(L)	R(V)	R(D)	R(L)	Justification (V)	Justification (D)	Justification (L)
16	C	C	C	R-K	R-K	R-K	Símptomes típics	Relació possible, tot i que el codi dels símptomes és molt inespecífic	Son signes que acaban en un diagnóstic de IC
17	I	C	P-C	N	R-K	R-K	Encara que són patologies que es relacionen no podem establir cap causa-efecte	Relació possible	Vinculació clara
18	C	C	C	R-K	R-K	R-K	És un tractament típic	Relació clara i coneguda	Vinculació clara
19	C	C	C	R-K	R-K	R-K	Símptomes típics	Relació clara i coneguda	Vinculació clara
20	C	C	C	R-K	R-K	R-K	És un tractament típic	Relació clara i coneguda	Vinculació clara
21	C	C	C	R-K	R-K	R-K	És un tractament típic	Relació possible	Vinculació clara
22	I	C	P-C	N	R-K	R-K	Encara que són patologies que es relacionen no podem establir cap causa-efecte	Relació possible	Vinculació clara
23	C	C	C	R-K	R-K	R-K	Símptomes típics	Relació possible	Vinculació clara
24	P-C	C	C	R-K	R-K	R-K	Es plausible que un pacient amb MPOC no diagnosticada faci una sobreinfecció respiratòria, però no és el diagnòstic esperat per la presència de febre	Relació clara i coneguda	Vinculació clara

ID	C(V)	C(D)	C(L)	R(V)	R(D)	R(L)	Justification (V)	Justification (D)	Justification (L)
25	P-C	C	C	R-K	R-K	R-K	Es plausible que un pacient amb MPOC no diagnosticada faci una sobreinfecció respiratòria, però no és el diagnòstic esperat per la presència de febre	Relació clara i coneguda	Vinculació clara

Table 31: Evaluation of the first pack results

E Evaluations Pack 2

The following table shows, the evaluation of each doctor for the second pack of rules. *C* represents the *Correctness*, *R* the *Relevance* of a rule. Regarding the letters inside the brackets, *L* refers to *Luis*, *D* to *David* and *V* to *Victor*. The ID of the rules correspond to the ones presented in Table 30. For visualization purposes, we also abbreviated the following labels:

- Correctness:
  - I: Totally incorrect.
  - C: Totally correct.
  - P-T: Partially correct (for the temporal aspect).
  - P-C: Partially correct (for the clinical aspect).
  - P-B: Partially correct (for both aspects).
- Relevancy:
  - N: Not relevant.
  - R-K: Relevant and known.
  - R-U: Relevant and unknown.

The results can be observed in the next page.

ID	Corr (V)	Corr (D)	Corr (L)	Rel (V)	Rel (D)	Rel (L)	Justification (V)	Justification (D)	Justification (L)
1	P-B	P-T	P-C	R-K	N	R-K	El diagnòstic és molt imprecís i el periode de temps molt llarg.	Símptoma massa inespecífic	L'infart cerebral pot anar precedit de diversos signes neurològics
2	C	P-T	I	R-K	N	N	És el tractament del diagnòstic	Relació lògica. Dx poc específic	La majoria de gent gran pren AINEs o analgèsics
3	C	P-T	I	R-K	R-K	N	És el tractament del diagnòstic	Relació lògica. Dx poc específic	La majoria de gent gran pren AINEs o analgèsics
4	C	I	P-C	N	N	R-K	Diagnòstics molt similars	Símptoma i dx iguals	La presencia d'untumor primari afavoria la de unt umor secundari
5	P-C	I	P-C	R-K	N	R-K	Hem d'entdre que una persona amb neoplàsia si te dolor es prescriuen aquest tipus de fàrmacs.	Relació no aporta res	Abans del seu diagnóstic el tumor pot provocar dolor
6	C	I	C	R-K	N	R-K	És un simptoma típic	Una flebitis no dóna dolor generlitzat	Es frequent que abans d'arribar al diagnóstic de tromboflebitis el pacient vingui per dolor a les cames
7	I	P-T	C	N	R-K	R-K	Es una confusió al condificar possiblement al posar dolor en algún lloc de les cames.	Possible relació entre dolor extremitat i flebitis. Massa temps entre antec i dx.	Els antecedents acostumen a precedir al diagnostic
8	I	I	C	N	N	R-K	Símptomes poc específics el de simptomas cardiovasculars	Una flebitis no dóna dolor generlitzat	Els antecedents acostumen a precedir al diagnostic
9	C	C	I	R-K	R-K	N	Tractament de l'anèmia	Ferro com a tto de l'anèmia	En tot cas al revés

ID	Corr (V)	Corr (D)	Corr (L)	Rel (V)	Rel (D)	Rel (L)	Justification (V)	Justification (D)	Justification (L)
10	C	C	C	R-K	R-K	R-K	És un simptoma típic	Possible símptoma d'anèmia	Es ben conegut
11	P-B	P-C	P-C	R-K	R-K	R-K	Encara que els 2 símptomes són inespecífics podien entendre's com manca d'arie, fatiga i també com a taquicàrdia que són simptomes de l'anèmia	Codis símptoma massa inespecífic. Suposo que fa referència a "dispnea"	Abans d'arribar al diagnostic es frequent la disnea
12	I	I	I	N	N	N		Potser fa referència a elevació TA secundària a anèmia?	Son independents, seria més correcte hipotensió
13	C	I	I	R-K	N	N	Moltes vegades una neolàsia de pulmó debuta amb una pneumònia que no es resol	En algun cas pneumonia pot ser indicatiu de neoplàsia, però no és raro	No veig cap relació
14	C	C	P-C	R-K	R-K	R-K	Són simptomes típics	Símptomes possibles	Relació entre canvi de ritme intestinal i neo de colon
15	C	C	C	R-K	R-K	R-K	És un simptoma típic	Símptomes possibles	Relació entre canvi de ritme intestinal i neo de colon
16	C	C	C	R-K	R-K	R-K	Són simptomes típics	Símptomes possibles	Tots poden ser signes previs a una neo
17	C	I	C	R-K	N	R-K	És el tractament del diagnòstic	És una relació massa inespecífica	Tots poden ser signes previs a una neo

ID	Corr (V)	Corr (D)	Corr (L)	Rel (V)	Rel (D)	Rel (L)	Justification (V)	Justification (D)	Justification (L)
18	I	C	P-C	N	R-K	R-K		Símptomes possibles	Tots poden ser signes previs a una neo
19	C	I	C	R-K	N	R-K	Són simptomes típics	És una relació massa inespecífica	Tots poden ser signes previs a una neo
20	C	C	C	R-K	R-K	R-K	Són simptomes típics	Símptomes possibles	Tots poden ser signes previs a una neo
21	C	C	C	R-K	R-K	R-K	Són simptomes típics	Símptomes possibles	Tots poden ser signes previs a una neo
22	P-C	C	C	R-K	R-K	R-K	Moltes vegades una neolàsia de pulmó debuta amb infecció respiratòria que no es resol	Símptomes possibles	Tots poden ser signes previs a una neo
23	P-B	C	P-B	R-K	R-K	R-U	Entenc l'antecedent com a pacient fumador, de manera que és un antecedent de neoplàsia, normalment precedeix en anys al desenvolupament de la neoplàsia	Símptomes possibles	El tabac causa neoplasia, però trastorns del comportament?????
24	I	I	P-C	N	N	R-K		Aquest tto no hauria de fer pensar en una neoplàsia, hi ha altres patologies més frq	La tos com a signes previ a neo de pulmó

ID	Corr (V)	Corr (D)	Corr (L)	Rel (V)	Rel (D)	Rel (L)	Justification (V)	Justification (D)	Justification (L)
25	C	C	P-C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures i antibiòtic per úlceres infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
26	C	C	P-C	R-K	R-K	R-K	Els pacients ATDOM pateixen caigudes i en analítiques poden presenta enzims musculars elevats	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
27	C	C	C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures i antibiòtic per úlceres infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
28	C	C	C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen antibiòtic per ferides infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
29	C	C	P-C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures i antibiòtic per ferides infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM

ID	Corr (V)	Corr (D)	Corr (L)	Rel (V)	Rel (D)	Rel (L)	Justification (V)	Justification (D)	Justification (L)
30	C	C	P-C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures i antibiòtic per ferides infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
31	C	C	P-C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures, antibiòtic per ferides infectades.	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
32	C	C	P-C	R-K	R-K	R-K	És fàcil i conegut que els pacients d'atenció domicilària requereixen cures i antibiòtic per ferides infectades	És habitual en pacients domiciliaris que precisin cures tòpiques de ferides	Tractament de nafres com pas previ a ATDOM
33	I	I	P-C	N	N	R-K		Codi símptoma massa inespecífic	Depen de la causa del dolor (metàstasi. . .)
34	P-B	I	I	R-K	N	N	L'eritema dels pacients amb atenció domiciliària és dermatitis del bolquer	Codi símptoma massa inespecífic	No es motiu de ATDOM, a no ser que sigui reacció anafilàctica

Table 32: Evaluation of the second pack results