POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Thesis

Methods for Removing Non-Interesting Itemsets when Mining Electronic Healthcare Records

Supervisor:
prof. Silvia Chiusano
Co-supervisor:
prof. Ricard Gavaldá

Candidate
Vincenzo Genna

Academic Year 2017-2018

Abstract

The usage of data mining techniques in healthcare has exponentially increased in the last years. Analyzing the huge amount of data that is nowadays produced by healthcare systems can lead to the extraction of useful and interesting informations about patients and diseases, which can be exploited to improve medical research and knowledge. Understanding how diseases and other characteristics of a patient are interrelated is a crucial point because it can help healthcare specialists to focus only on important factors when addressing cures for a given clinical case. Frequent itemset mining techniques are widely used for this purpose, but they can lead to the retrieval of too many redundant or not interesting pieces of information. In this project we study and report performances of a measure proposed to remove redundant and irrelevant rules from data and suggest an approach to unveil the main comorbidities for a given disease, along with the possibility to use the latter results to further filter not interesting informations. Results show the effectiveness of the two studied methods as also proved by the main literature that was reviewed during the project, even if we suggest the collaboration with healthcare specialists in order to get more relevant outcomes.

Keywords: healthcare analytics; itemset mining; chronic disease; interestingness measures; suppport to clinical decision making; comorbidities analysis.

Acknowledgements

My sincere thanks to my supervisors for their constant help along the duration of the project.
My dad, my mum and my brothers, I cannot ever thank them enough for supporting me on every single step I have made.
I would like to acknowledge all the guys that have been close to me since high school in Marsala and the guys with whom I have been studying for (tons of) exams in Turin.
Thanks to the Paraaaaa, laughing and joking with me about everything during the last five years.
I am also grateful to all the people I have met in Barcelona during the fastest year of my life, who were always ready to help me when I needed.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Chronic diseases affect health and quality of life. These are diseases that can affect a patient's life for at least 3 months, but that can easily persist for many years. A disease is usually defined as chronic if it cannot be cured by current therapies and if its effects last for a long-term period. The most prevalent chronic diseases in developed countries are heart diseases, stroke, cancer, type 2 diabetes, obesity, and arthritis. Apart from being linked to more than half of the deaths that occur each year, especially in economically developed countries, chronic diseases, and in particular their related cures, are also responsible for the majority of the healthcare costs, up to 70% according to estimations. This probably happens because the real causes of chronic diseases are not well understood, so the treatments for these diseases are superficial or stopgaps and never getting to the actual basis of these diseases. Moreover, most of the times chronic diseases are present along with other diseases (comorbidities), often in combinations that can turn out to be quite complex and that can also lead to further cost increments if the relationships between them are unknown, or not taken into account.

Data mining techniques and methodologies can, in this regard, be used to analyze the data produced by healthcare transactions, that are nowadays too complex and voluminous to be processed and analyzed by traditional methods. To look into these combinations may find out

not obvious links and associations between diseases, that can be further explored by healthcare specialists' studies to figure out how to improve cures (and their related costs) for certain diseases.

1.2 Context

Healthcare industry today produces huge amounts of different data about hospitals, diagnosis, patients. Processing these data for knowledge extraction is a fundamental task because that can give support for understanding complex informations in healthcare industry.

In last decade, there has been a boost in usage of data mining techniques on medical data to determine useful trends or patterns that are used by health-care specialists in disease analysis and decision making when it comes to diagnosis or drugs prescription. According to a survey published by *PubMed* [30], data mining is becoming increasingly popular in health-care, if not increasingly essential.

Pattern mining is a data mining technique used to identify patterns such as trees or sequences or sets of items within a database of transaction. It has in particular been used to examine EHR, due to the huge availability of transactional records, where each patient is associated to a set of clinical information. Frequent patterns are itemsets or subsequences that appear in a data set with frequency (defined as *support*) higher than a given threshold. Being able to find this kind of patterns is crucial to mine associations, correlations or many other interesting relationships among data. This is one of the reasons why frequent pattern mining or frequent itemset mining have become widely used to analyze EHR.

One of the known issues of frequent itemset mining is the huge amount of itemsets that are mined, many of which are not interesting, redundant or well known to the domain experts, so a filter is needed. Apart from the *support* measure, different interestingness evaluation criteria have been defined, either based only on the raw data (relying on theories in probability, statistics, or information theory) or taking into account both the data and the user of these data (using, if available, the user's domain or background knowledge about the data). Some

of these measures will be used in order to address the problem of redundancy, that appears when two items A and B show an high association, but this association is not only dependent on A and B, but on their stronger association to a common item C. In this perspective, the theoretical concepts studied in a previous master thesis [43], also developed in UPC, will also be used in our analysis to propose and evaluate an additional interestingness criterion and its performances.

Studies have been conducted, applying frequent itemset mining, on how the treatment of a disease can affect the co-existing other disorders (comorbidities) [31], or to identify temporal relationships between medications and accurately predict the next medication likely to be prescribed for a patient [32]. Other ones have analyzed the comorbidities of diabetes filtering results by considering contextual information (patient demographics, treatment, status) about extracted patterns [33] or have used association rule mining for the same purpose [34]. Another approach to study comorbidities will be developed in this work, in order to rank the most important ones for a given diagnostic and to possibly help in further filtering of redundant and not interesting associations.

This work has been carried out as part of a project in collaboration with the LARCA research group of Universidad Politécnica de Catalunya, whose research line analyzes Electronic Healthcare Records (EHR) by means of data mining and machine learning algorithms. LARCA cooperates with several health care institutions in Catalonia such as the Servei Català de la Salut (CatSalut), Institut Català d'Oncologia, and Hospital de Sant Pau, who provide data to be analyzed. In our project we analyze data from Hospital de Sant Pau.

The research group is currently working on the development of a software tool that should help doctors in the search for interesting associations between diagnostics. The remarks and observations of the specialists further confirmed the need to implement a tool able to remove redundancy and other not interesting informations from the large quantity of output results that are otherwise produced.

1.3 Objectives

In order to remove or filter out from the analysis some of the many associations that are mined, we use the *metalift* measure approach to detect not interesting itemsets and evaluate its performances over different studied populations.

After removing the not interesting associations, we analyze the explanatory power of factors that are strongly linked to the diagnostic of interest (e.g. comorbidities of a disease) on the remaining itemsets. This approach allows us to:

- detect associations that are not directly affected by the diagnostics we are studying, but that are explained by some factors that are related to them, and choose whether further filtering them or studying them in depth;

- compare the effects of comorbidities and other factors (sex, age) on a diagnostic of interest and rank the most important ones among these.

1.4 Structure of the report

This report is strucured as described as follows. This chapter has presented the motivation and context of the study and pointed out the main goals we want to achieve.

Chapter 2 explains the main background notions about the healthcare domain and frequent itemset mining, focusing also on some statistical measure of interest.

In *Chapter 3*, we report the results we found by reviewing literature about frequent itemset mining and data mining approaches applied to healthcare problems.

Chapter 4 is used for a descriptive analysis of the data we studied and to outline the methods we adopted to preprocess them in order to apply a standard itemset mining algorithm.

Chapter 5 outlines the main methods we developed and applied on the data we obtained after the steps in chapter 4 along with their formal definitions.

The main results and examples we found during the analysis are showed in *Chapter 6* and *Chapter 7*, along with some charts and graphics.

The last chapter highlights the main achievements of this thesis and suggests some future applications that can benefit from this project.

Chapter 2

Background

In this chapter we will give the necessary background on how healtcare systems and facilities are collecting their data and on the basics of frequent itemsets mining.

2.1 Electronic Healthcare

As explained in section 1.1, storing healthcare data in electronic records (EHR) can improve healthcare quality while reducing its costs. Data mining techniques can be used and have been used to retrieve useful and interesting relationships from the large amount of available EHR data nowadays.

throughout . These codes are organized in a hierarchical model, in the format ddd.d or ddd.dd where ddd specifies the general disease and the .d (or .dd) specifies the subgroup of the ddd disease. For example, the code 250 identifies diabetes mellitus while the code 250.1 is related to diabetes with ketoacidosis and 250.2 to diabetes with hyperosmolarity, while 250.21 encodes diabetes with hyperosmolarity type 1. Apart from numerical codes there are also the so called *V*-codes, represented in the format of *V*dd.d, added to deal with encounters for circumstances other than a disease or injury, that are useful to classify patient anamnesis. They represent a circumstance or past problem that is not the main reason of the current patient visit,

but that the doctor should be aware of. For example, *V*15.82 indicates a history of tobacco use, while *V*10.03 indicates that this patient also suffers or has suffered esophagus cancer - although today he came for some other problem.

The ICD-9-CM coding system is now moving to a new coding set, ICD-10-CM. This newer revision's list of codes contains more than 68,000 diagnostic codes, compared to 13,000 in ICD-9-CM. ICD-10-CM codes include also twice as many categories [35].

The majority of healthcare facilities started using ICD-10 codes at the end of 2015. Since the data we were provided were gathered from 2014, ICD-9 standard was kept for this project.

2.2 Itemset mining

2.2.1 Basic Notions

A transaction is typically the description of an event, or, in the healthcare field, of a patient admission, that is represented by a unique identification number (TID) and a list of the items making up the transaction (for example, diseases found to be present in the patient at the moment of the admission). A transactional database consists of a file where each record represents a transaction In the cases described in this project, we will refer to the transactional database as the *population* we are working on. Within this project, an item is one particular disease encoded by means of ICD-9 system or another information regarding personal attributes of a given patient, such as age or sex. An *itemset* is a set composed of zero, one, or more items among the ones present in the database. A *k-itemset* is an itemset of cardinality k.

2.2.2 Support and Lift

The *support* of an itemset is a measure that represents its frequency in the dataset. Let I be an itemset and $U = \{t \in T \mid I \subseteq t\}$ where T is the list of all transactions and $supp_{abs}(I) = |U|$. The relative support of the itemset I simply denoted as $supp(I)$ represent the proportion of

the transactions that contains the itemset I among all the transaction in the dataset and it is defined as $supp(I) = |U|/|T|$ where $|U|$ and $|T|$ are the number of transactions contained in U and T.

The *lift* of an itemset is one of the most used interestingness measures in itemset mining and pattern mining. In particular, given an itemset $I = \{A, B\}$, $lift(I)$ quantifies how much the two items A and B are related to each other, by comparing the probability of finding both together to the probability of them being completely independent. Formally:

$$lift(\{A, B\}) = \frac{supp(A \cup B)}{supp(A) * supp(B)}$$

Its value can range from 0 to infinity. A value close to 1 shows that the the probability of finding both items together in a transaction is really similar to the probability under the independence assumption, so the two items should be considered uncorrelated and the itemset not interesting. A lift greater than one, instead, indicates positive correlation between the two items, meaning that the occurrence of item A increases (or *lifts*) the probability of B to be in the same transaction. On the contrary, a lift value less than one shows that A and B are negatively correlated and that the presence of A is linked to the absence of B.

So far, only the notion of *lift* for 2-itemsets has been explained. There could be many reasonable definitions of *lift* for larger itemsets, in this project we decided to define it as the maximum among the lifts of any possible partition in two subsets:

$$lift(I) = min_{a \in I} \left\{ \frac{supp(I)}{supp(a) * supp(I \setminus \{a\})} \right\}$$

This definition allows us to continue to look at itemsets with high lift as the ones that adds more informations on the relationships between the items they contain.

Other definitions in the literature include taking the minimum over all partitions of I into two subsets. This increases exponentially the number of checks to perform, and it is not clear in practice that it adds much to our chosen definition. All our methods could be extended to this other definition if desired.

2.2.3 Risk ratio

Risk ratio is a statistic method that is often chosen to analyze binary events by looking at binary factors, especially in cases where the probability of the event of interest is low. Having a given *exposure* factor and an *outcome* of interest that has to be studied, risk ratio is the ratio of the probability of the *outcome* occurring in the population where the *exposure* is true over the probability of the same *outcome* occurring in the non-exposed population.

Given the *contingency table* shown in Table 2.1, the *Risk Ratio* for an exposure e and an outcome o is defined as

$$RR(e, o) = \frac{A/(A + B)}{C/(C + D)}$$

An $RR \approx 1$ means that the exposure factor is not related to the presence of the outcome. An $RR < 1$ means the outcome is less likely to be present in the population where the exposure is present than in the one where the exposure is not present, that is the the presence of the outcome is linked to the absence of the exposure. Backwards, $RR > 1$ says the presence of the outcome is more probable when the exposure factor is present.

In this project, we will refer to the *exposure* as a binary factor x which the base population can be split on, as an attribute for age or sex or a certain diagnostic. The *outcome*, instead, will always be referred as the presence or the non-presence of an itemset I. So, in the cases studied in this project, $RR(x, I)$ will basically be the ratio of the support of I on the population having x on the support of I on the remaining population.

One of the advantages of the risk ratio is that the *statistical significance* of its value can be evaluated by means of *confidence intervals*. Confidence in a value is usually higher if the noise (intended as values that are present in the analysis, but that are actually useless) is lower and/or the sample size is larger.

In this project confidence intervals were calculated from $RR(x, I)$ value referring to the formula proposed by [44]. Also a *p-value* can be derived given a confidence interval.

Table 2.1: Contingency table for risk ratio

	Outcome	$\overline{Outcome}$
Exposure	A	B
$\overline{Exposure}$	C	D

2.2.4 Algorithms

As mentioned before, we will focus on the branch of pattern mining involved in the discovery of interesting itemsets among all the transactions.

An itemset is said to be *frequent* if its support is greater than or equal to a given threshold.

Frequent itemset mining is most of the times explained by its usage for market basket analysis. Market basket analysis is performed using algorithms to identify associations, or patterns (itemsets), among the various items that have been chosen by a particular customer and placed in their market basket during the same shopping session. Several itemset associations are usually found (milk, bread and beer, diapers are two of the most common/famous examples), even in small datasets. Support and lift are generally exploited as a preliminary filter to remove some of them.

Since the first algorithm was developed in 1993 [1], several other implementations and enhancement have been proposed. Looking for all the possible combinations of items requires several database scans, so all the most famous implementations focus more on improving memory and CPU usage, and less on dealing with the huge amount of frequent itemsets that are mined, especially when the minimum support threshold is low.

Usually, an end user is interested in a small subset of all the patterns that are mined, but, due to the abstract nature of interestingness, there is no common agreement on a formal definition of it in this context.

Most of the time *interestingness* is more linked to domain-specific knowledge than to mere data properties. For this reason, many other interestingness measures have been studied and applied to real data.

Different algorithms and interestingness measures proposals will be described in the next chapter.

Chapter 3

State of The Art

3.1 Frequent Itemset Mining

The topic of frequent itemset mining has experienced an increasing interest from the data mining research community towards it in the last 25 years. This has led to great progresses in this field, bringing itemset mining algorithms from being used for applications similar to market basket analysis to become nowadays used also in complex machine learning tasks such as time series classification and clustering or leading to the development of deep learning models to solve pattern recognition problems.

The idea of mining frequent itemsets from data was first proposed by Agrawal et al., who developed Apriori, an algorithm to generate all the significant association rules between items in a transactional database [1]. The algorithm makes many searches in database to find frequent itemsets where k-itemsets are used to generate $(k+1)$-itemsets. Each k-itemset support must be greater than or equal to a minimum support threshold to be considered as frequent. It scans the whole database to find supports of 1-itemsets. The support of the 1-itemsets is used to find the 2-itemsets and so on until there are no more k-itemsets for the considered k. The search space is pruned by means of the Apriori principle [2], stating that if a k-itemset is not frequent, none of its $(k+1)$-superset will be either.

Apriori algorithm's main weakness is that it is costly in terms of time to explore a huge number of candidate sets that appears if a low minimum support threshold is set or if the frequent itemsets' cardinality becomes too large. For example, if there are 10^4 frequent 1-itemsets, the algorithm will generate more than 10^7 2-itemsets as candidates [3]. Moreover, it might potentially generate 2^{100} candidate itemsets to detect frequent pattern composed by 100 items [4] (although that rarely happens in practice). Apart from potentially checking a large number of candidates, it will also scan the database many times repeatedly to find candidate itemsets, once per every k. Then, Apriori can turn out to be very low and inefficient, especially when memory and CPU resources are limited. To overcome these limitations, several extensions have been proposed such as hashing techniques [5], sampling approach [6], dynamic itemset counting [7], incremental mining [8], parallel and distributed mining [5, 9, 8, 10].

The FP-Growth approach by [11] was developed to eliminate some of the bottlenecks in Apriori, candidate generation and several database scans. It uses a structure called an FP-Tree. In an FP-Tree each node is used to represent an item and its support, and each branch represents an association between two nodes it links. The biggest benefit of FP-Growth is that the algorithm only needs to read the file twice, avoiding database scan for every iteration like Apriori. Another huge advantage is that it removes the need to calculate the pairs to be counted, which is very processing heavy, because it uses the FP-Tree. This makes it $\mathcal{O}(n)$ which is much faster than Apriori [12]. The FP-Growth algorithm keeps in memory a compact representation of the database.

Several extension of FP-growth have been published. [13] proposed a depth-first generation of frequent itemsets, an hyper-structure mining of frequent patterns approach was carried out by [14] using recursive exploration of the structure and top-down and bottom-up traversal method was developed by [15].

Apart from being mined in horizontal format transactions (where each transaction is associated to its Transaction ID), itemsets can also be mined in vertical format transactions (where each item is associated to the list of the transactions it appears in). Algorithms based on vertical format mining are among the most common algorithms currently being used and researched

because they implement fast support counting and automatic pruning of irrelevant data.

One of the most famous vertical format mining algorithm is ECLAT [16]. It requires only two scans of the database, similar to FP-Growth, the first to eliminate items that are found to be infrequent, and the second, to load the transactions in vertical format.

Most of the existing vertical format based algorithms are modifications of ECLAT, [17] showed that some data mining problems can also be solved using the general purpose database management systems (DBMS) and storing data in vertical format with quite impressive results.

One of the biggest issues when mining frequent itemsets from real (large) datasets is the genaration of a huge number of itemsets, especially when the minimum support threshold is low. This happens because, as explained by the Apriori principle, if an itemset is found to be frequent, each of its subsets is frequent. To address this problem, closed frequent itemset mining [18] was developed, using an Apriori-based algorithm called A-Close. An itemset A is a closed frequent itemset in a dataset if A is frequent in the dataset and a superset of A with A's same support does not exist.

For the same reason maximal frequent itemset mining [19] was also proposed, stating that an itemset A is a maximal frequent itemset in a dataset if A is frequent all its supersets are not frequent.

Other important work on closed itemset mining include studies like CLOSET [20], FPClose [21], CLOSET+ [22], CHARM [23].

Apart from trying to remove the hundreds of frequent itemsets that are produced by the algorithms mentioned above, a good itemset mining approach should distinguish the most useful patterns from those that are obvious or are already well known to the domain specialists. It is necessary to filter out those patterns through the use of some measure of "usefulness" or "interestingness".

Different works has been carried out to determine what is interesting, with an agreement that interestingness is basically subjective. While there is no single definition of interestingness

yet, the shared idea among experts is that finding interesting rules is a very difficult problem, requiring domain knowledge and/or user interaction.

As [24] explains well, interestingness measures are divided into objective measures, that rely on the statistical properties of the discovered itemsets, and subjective measures which exploit the users knowledge of their particular problem domain. Classical objective interestingness measures are the lift (in [25] defined as *interest*) and the support, that were explained in subsection 2.2.1, but the total amount is quite huge and includes measures such as *Jaccard index* or the *cosine*, both used to determine similarities [26] [27].

An extensive review of 21 measures was conducted by [25] who examined objective interestingness measures, comparing them using several properties and showing that the validity and performance of an interestingness measure, often depends on the domain of the data which the measure is used for and that there is no measure that is better than the others in all cases.

As for subjective measures, we will cite two of them: The one proposed in [28] exploits user's domain knowledge to classify simple rules to quickly eliminate the non interesting rules and to construct a domain knowledge base. And the work in [29], with a Bayesian approach to determine when a rule can be classified as "unexpected", by relying on a set of rules that the end user marked as true or not.

Though the majority of interestingness measures were developed in order to be applied to evaluate association rules mining, the concepts can be extended to the more general task of frequent itemset mining.

3.2 Data Mining for Healthcare

As mentioned before in this work, last decade has witnessed a boost in usage of data mining techniques on healthcare data in order to discover patterns that are used by specialists in disease analysis and decision making when it comes to diagnosis or drugs prescription. As stated by [30], data mining is becoming increasingly popular in health-care, if not increasingly essential.

Studies like [32] have been conducted to find out that sequential pattern mining is an effective technique to identify temporal relationships between medications and can be used to predict next steps in a patients medication regimen and that accurate predictions can be made without using the patients entire medication history. Frequent sequence mining (a branch of frequent pattern mining where also the order in which the items appear is important, and is taken into account) was applied in a study that aimed to analyze the effect that the treatment for a given disease also affects its co-existing other disorders (comorbidities) [31].

One study carried out by [36] aimed to predict patients who are at risk of developing heparin-induced thrombocytopenia and presented a framework to generate a small set of predictive and relevant patterns, filtering those who were considered not useful to the classification task.

Maximal frequent itemsets was applied to outpatient records gathered by the Bulgarian National Health Insurance in 2010-2016 for more than 5 million citizens yearly in order to analyze the comorbidities of diabetes, schizophrenia and hyperprolactinemia and then filtering results by considering contextual information (patient demographics, treatment, status) about extracted itemsets [33] while [34] used association rule mining to analyze comorbidity in patients with type 2 diabetes mellitus. A comorbidity network for different types of cancer was constructed by[37] by applying large-scale itemset mining approach among millions of patients and used this to show associations between cancers and their comorbidity relationships with various kinds of diseases.

Chapter 4

Dataset and Data Preprocessing

4.1 Dataset

The data used in this project were provided by Hospital Sant Pau, Barcelona, under an agreement with UPC (and its data science research group, LARCA).

The dataset is a collection of Electronic Health-care Records, gathered between 2014 and 2016 and stored in CSV format, where each row contains informations related to a single hospital visit or admission. Each record reports basic informations on the patient (anonymized id, date of birth, sex) and on the hospital episode (date of admission, date of release, urgency, priority, reason of release) along with the detailed descriptions of the diagnostics and the procedures related to the episode. The diagnostics and procedure informations are classified in several columns, from the main diagnostic, labeled with column name *DP*, to other secondary diagnostics (up to 14 of them, labeled with column names from *DS1* to *DS14*) and from the main procedure, if any, applied to the patient, labeled with column name *PP*, to secondary procedures (up to 10 of them, labeled with column names from *PS1* to *PS10*).

Within this project, columns related to procedures were discarded and only the basic informations (excluding the anonymized patient ID, considered not relevant for the analysis) and the diagnostics ones were kept. Concerning the diagnostics, the main diagnostic *DP* might be

considered more important than the secondary diagnostics *DSi*, but the order within secondary diagnostics is generally considered not relevant.

Diseases and other health-care informations were encoded by means of the ICD-9-CM standard, described in section 2.1. In the original dataset, and additional column was present for each of the ICD-9-CM encoded columns, to provide a textual description of the code. These columns were removed for the analysis.

The dataset contains information for 79533 episodes, and only 18 columns (sex, admission circumstances, and the 15 ones related to diagnostics) were kept out of the original 60 column.

Table 4.1 shows the distribution of the patients in our dataset by age group and sex in terms of absolute value (and in terms of percentage value over the whole dataset), while figure 4.1 compares the supports of the most prevalent diagnostics in the dataset. The most prevalent diagnostics for different population subsets are listed in appendix B. *Essential hypertension* (401.9), *Hyperlipidemia* (272.4), *Atrial fibrillation* (427.31) and *Diabetes mellitus* (250.00) are the most common diagnostics, especially the first two of them.

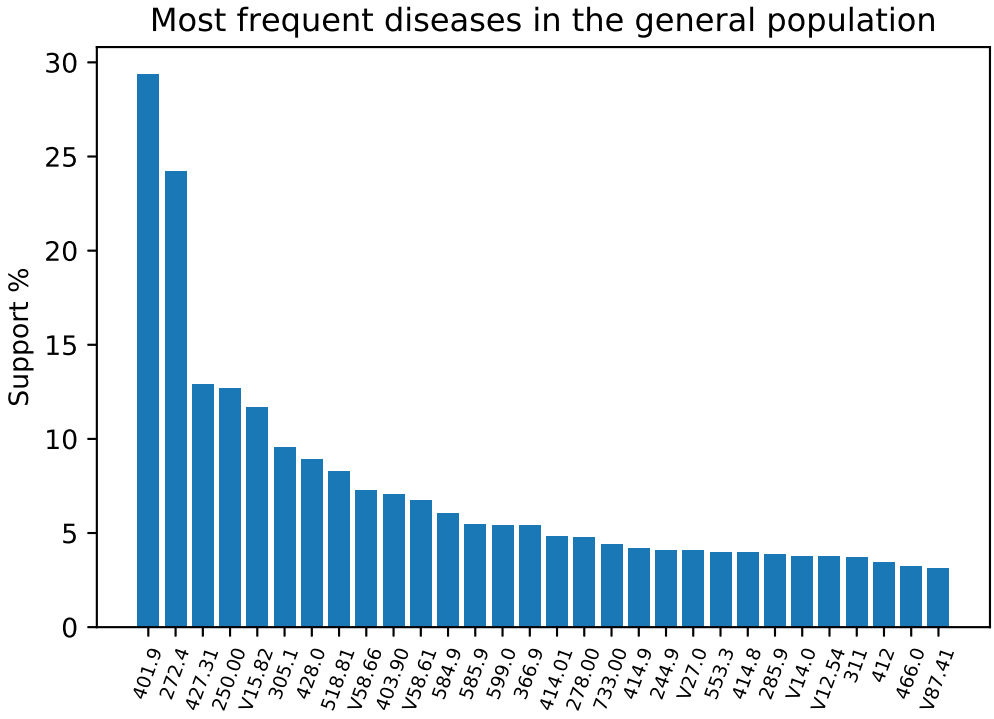


Figure 4.1: Most prevalent diagnostics

Table 4.1: Absolute (%) distribution of the patients by age group and sex

	M	F	
<40	5728 (7.2%)	8896 (11.18%)	14624 (18.39%)
≥ 40, <65	10479 (13.17%)	9929 (12.48%)	20408 (25.66%)
≥ 65, <85	16293 (20.48%)	15573 (19.58%)	31866 (40.06%)
≥ 85	4545 (5.71%)	8090 (10.17%)	12635 (15.89%)
	37045 (46.59%)	42488 (53.41%)	79533 (100%)

4.2 Data preprocessing

The important columns described in section 4.1 were extracted from the original data and an additional column reporting the age of the patient at the moment of the admission was computed for the analysis.

To be included in a transactional database, all the continuous attributes need to be discretized using some criteria.

In this case, the computed age was the attribute to discretize. The standard age ranges suggested by doctors doctors are 0-39, 40-64, 65-84, >85, so age should be replaced with four distinct binary attributes. However, in our analysis, we decided to encode age in another way, cumulatively, with 6 items: ≥ 40, ≥ 65, ≥ 85, and <40, <65, <85. The last three give the possibility of having itemsets that apply to young people. With this encoding method, age is translated into a set of three items, depending on its continuous value.

The ICD-9-CM codes that appeared to have a support lower then a specified threshold were excluded from analysis, because diseases that were present only in a small percentage of the whole population were considered not useful. Figure 4.2 shows the amount of diagnostics that are left to analyze in the dataset after setting the said threshold to several values. It is possible to notice that even with a low threshold (0.1%) the number of diagnostics that will be analyzed reduces to 12.81% and that this threshold, on this dataset, has a significant impact until it is set to about 1%.

The given data were converted into transactional format, by describing each patient with a list of the diseases he/she had at the moment of the episode and with the categorical informa-

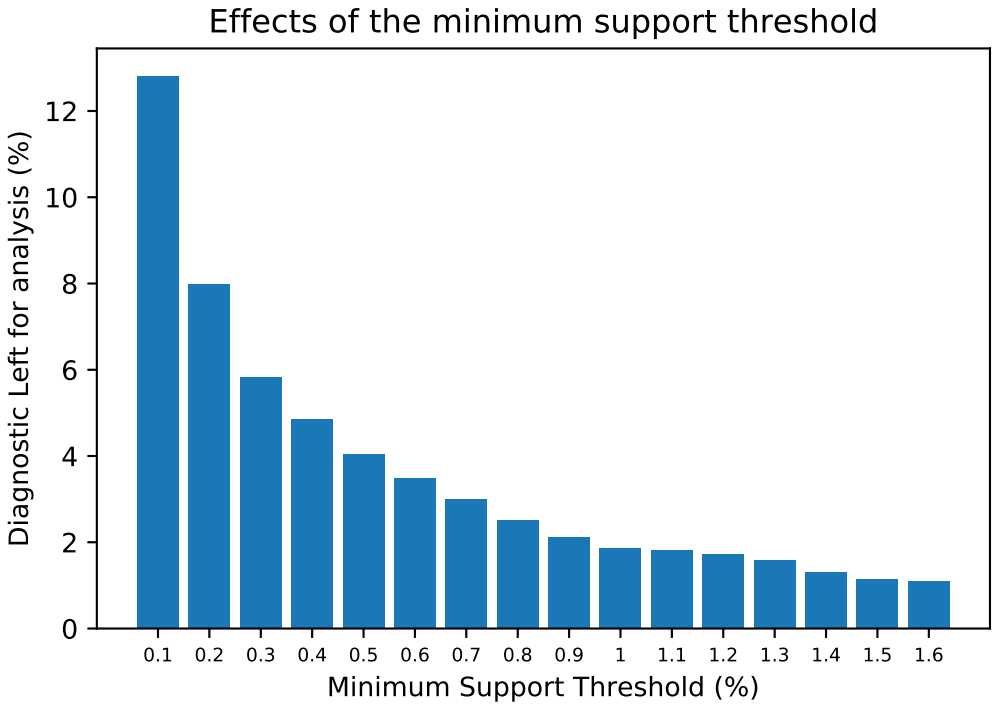


Figure 4.2: Effects of the support threshold on the diagnostics to analyze

tions referring to age and sex.

4.3 Itemset mining

The original data, after being preprocessed as described above, were then used as input for the Equivalence Class Transformation (ECLAT) algorithm [10], which we used to retrieve frequent itemsets from our data. The implementation we used [38] allowed us to perform an efficient and fast mining, giving also the possibility of choosing among different settings and parameters.

After selecting a minimum support threshold of 0.25% and without filtering by lift (to avoid cases in which an itemset is not important in the general population, but it becomes interesting when other factors are considered), we obtained 4510 frequent itemset, along with their associated informations about support and lift measures.

Chapter 5

Metalift and Explanatory Variables

5.1 Metalift measure

Metalift is a measure that was studied in a previous project carried out at UPC [43] as mentioned in section 1.2. Given a base population A (the population we extracted after preprocessing the original data), we can select a proper subpopulation B by requiring the presence of one item x in all of B's transactions. So we can formally define $B = \{a \in A | x \in a\}$. For an itemset I that, after itemset mining, turns out to be frequent on both populations A and B, $metalift(I, A, B)$ is formally defined as follow:

$$metalift(I, A, B) = \frac{lift_B(I)}{lift_A(I)}$$

Its value can range from 0 to infinite. In the cases described in this project, specific diseases or age and sex attributes will be used as items x to define subpopulations. By requiring the presence of x in the subset B, we aim to exclude its effects on the frequent itemsets we found, to eventually check the importance it has in making those association stronger or weaker.

A $metalift < 1$ indicates that the items in the itemset I are less related in subpopulation B than what they were in the general population, because their association depends somehow on the x factor. While a $metalift > 1$ suggests that the items in I show an increased association

between them when the item x is present.

Hence, these itemsets that show significant changes in the $lift$ value may be classificated as important or interesting by doctors or healthcare specialists because these cases add informations about the effects of the x factor on the diseases contained in the itemset. But, in section 6.1, we also described similar cases that can be marked as not interesting or obvious by specialists, so, in general, these are associations that require further studies or opinions by healthcare insiders.

When $metalift \approx 1$, instead, the itemset shows no major changes in term of lift value between its components, so the x factor can be considered as a non influent factor for the association among them and the given itemset can be filtered from the analysis.

5.2 Filtering by metalift

We applied this filtering method as a preliminary step before further analyzing the itemsets, introducing a parameter $\epsilon_{metalift}$ to set the range of $metalift$ values to filter. In particular we used this approach, for different diagnostics x, on all the itemsets that were found to be frequent, after applying the algorithm described in section 4.3, both in the general population A and in the subpopulation B, filtering all the itemset I where $|1 - metalift(I, A, B)| \leq \epsilon_{metalift}$.

5.3 Explanatory variables

Another method we used to analyze lift variations and try to understand relationships between itemsets was to also look at variations in the support of the itemsets, with respect to other factors linked to the selected x disease a user is interested in. For example, as said in section 5.1, cases where $metalift < 1$ may represent situations where the given itemset I seems to be related to x so that, when the effects of x are removed, the associations between items in I disappears too.

An hypothetical example that describes well the situation we are going to address is represented by the association between the diagnotics in the itemset

$$I = \{\textit{Breast Cancer, Having had child}\}.$$

In a population of 2000 individuals, half males and half females, we notice that 10 of them ($supp(I) = 0.005$) show the presence of both diagnostics in I and that 100 show one of them only ($supp(\textit{Breast cancer}) = supp(\textit{Having had a child}) = 0.05$). The resultant lift on the overall population is thus $lift(I) = 2$, that shows a significant association between the two items. However, if we consider only the female population, we will notice that the absolute supports remained unchanged, but, since the number if individuals is now 1000, $supp_F(I) = 0.01$ and $supp(\textit{Breast cancer}) = supp(\textit{Having had a child}) = 0.1$. This lead to $lift_F(I) = 1$, showing that the items are not associated anymore in the female population. This means that the apparently high lift may be almost totally explained by Sex, so reporting it as a pattern is misleading, as it says nothing interesting beyond the fact that females have can both children and breast cancer. So if we report the associations, that are trivial from a healthcare-knowledge point of view, {*Sex=F, Having had a child*} and {*Sex=F, Breast cancer*} it is best not to report this itemset I.

The lift reduction may be directly caused by a factor x chosen to select a subpopulation, as it's showed in the example above. But it can also be caused by another factor y that is strongly related to x, so that when the effects of x are removed, also y ones are eliminated, as explained in section 7.1 (*Example 2*).

In both cases we say that the association between elements of the itemset I is *explained by* y or that y is an *explanatory variable* for the itemset in the population affected by x .

Formally, we say that, in a population affected by disease d, x is an *explanatory variable* for the association of items in the itemset I if x is linked with d and $RR(x, I) > 1$ with a statistically significant confidence interval (or *p-value*). Since the itemsets that are analyzed when a disease d is selected are only the ones which d has a major impact on ($metalift \not\approx 1$), $RR(x, I) > 1$ would mean that I is rare on people without x and common on people with x

and this could help explain the association between the items in I. Being a factor that has to be used to filter a population, x can be any binary variable present in the dataset, from a diagnostic to age (encoded as a set of 3 binary variables as explained in section 4.2) and sex (M or F).

These informations need to be checked by healthcare domain specialists, because they can be trivial like the two examples cited above, which associations can be explained by *diabetes* and by sex, thus they could be further filtered. But there can be cases in which these informations may be actually interesting like the one showed in section 7.1 (*Example 3*).

Of course, also the problem of choosing which factors have to be analyzed for a given disease x is left to healthcare specialists or to reviews of the literature regarding x. Different ones can also be chosen by looking at the most frequent diagnostics (comorbidities) or attributes that appear with x. The last two approaches, checking comorbidities by reviewing literature and looking for most frequent factors, were the ones that we used in our different analyses.

5.4 Explanatory power

Using the method described above, we define *e*xplanatory power of a factor x for a disease d as the number of itemsets I which x is an *e*xplanatory variable for. This let us create a rank of the diseases with most explanatory power for a given disease, so that the most important comorbidities for a given disease can be highlighted.

Apart from this rank of diseases, age or sex were added to the rank even if their influence was calculated, by means of the same method described in section 5.3, separately, because as one might expect these are two factors that highly and mostly influence all the association between the analyzed diagnostics.

5.5 Software Implementation

We implemented the different approaches described above in a software tool that provides the final user with the most important findings of the analysis. Our implementation, after preprocessing the data as explained in section 4.2, initially performs a frequent itemset mining on the original data by applying the concepts described in section 4.3.

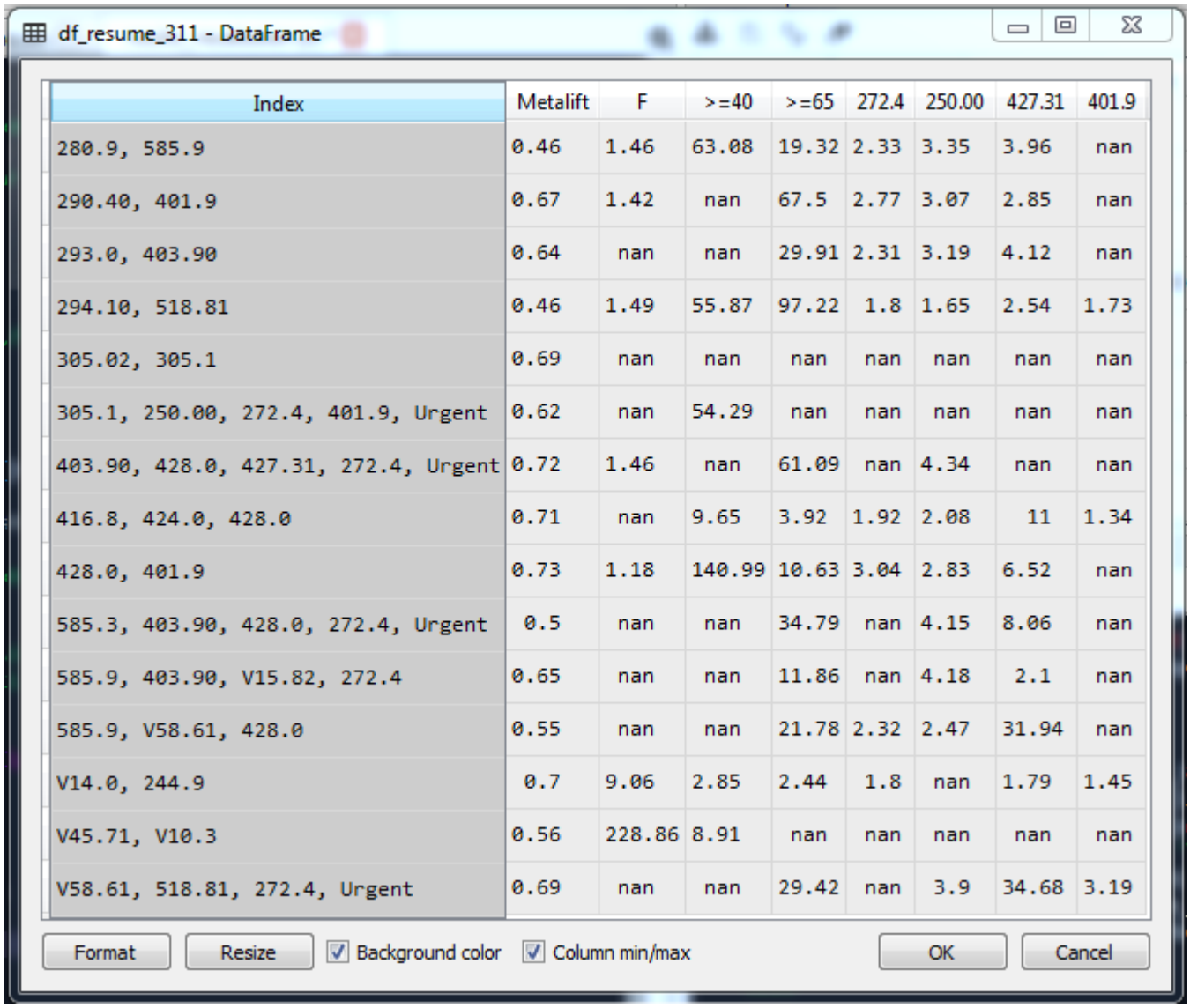
The users can then choose an interesting subpopulation by applying filters on sex or on age ranges or selecting one or more diseases the cases in the subpopulation should be affected by. The mining algorithm is applied on the subpopulations and its most frequent patterns are extracted.

The two set of frequent itemsets, the one mined from the original population and the one mined from the selected subpopulation, are then compared applying the *metalift* measure as described in section 5.1 and section 5.2, in order to filter out the itemsets that are not showing major changes between the two population (i.e. the ones that are potentially not interesting, according to the user's preferences).

The itemsets that are left are the ones that will be analyzed with the techiques described in section 5.3 and section 5.4.

In particular the users are asked to select a set of diagnostics that they want to study the explanatory role for. Age and sex are factors that are automatically included into the analysis if they are not selected, because they are considered the most important ones. For each itemset to analyze, the *Risk Ratio* of each potential explanatory variable is calculated, to discover how each of the factors included in the users-selected set affect the presence of the itemset. The factors that show an higher significant risk ratio are the ones that may be more important in explaining one association. The explanatory power of each one of this factors for the subpopulation the users are interested in is then computed.

An example of the final output of the software we implemented is shown in figure Figure 5.1. The analyzed subpopulation of interest is the one suffering from *Depressive disorder* (ICD9 code



Index	Metalift	F	>=40	>=65	272.4	250.00	427.31	401.9
280.9, 585.9	0.46	1.46	63.08	19.32	2.33	3.35	3.96	nan
290.40, 401.9	0.67	1.42	nan	67.5	2.77	3.07	2.85	nan
293.0, 403.90	0.64	nan	nan	29.91	2.31	3.19	4.12	nan
294.10, 518.81	0.46	1.49	55.87	97.22	1.8	1.65	2.54	1.73
305.02, 305.1	0.69	nan	nan	nan	nan	nan	nan	nan
305.1, 250.00, 272.4, 401.9, Urgent	0.62	nan	54.29	nan	nan	nan	nan	nan
403.90, 428.0, 427.31, 272.4, Urgent	0.72	1.46	nan	61.09	nan	4.34	nan	nan
416.8, 424.0, 428.0	0.71	nan	9.65	3.92	1.92	2.08	11	1.34
428.0, 401.9	0.73	1.18	140.99	10.63	3.04	2.83	6.52	nan
585.3, 403.90, 428.0, 272.4, Urgent	0.5	nan	nan	34.79	nan	4.15	8.06	nan
585.9, 403.90, V15.82, 272.4	0.65	nan	nan	11.86	nan	4.18	2.1	nan
585.9, V58.61, 428.0	0.55	nan	nan	21.78	2.32	2.47	31.94	nan
V14.0, 244.9	0.7	9.06	2.85	2.44	1.8	nan	1.79	1.45
V45.71, V10.3	0.56	228.86	8.91	nan	nan	nan	nan	nan
V58.61, 518.81, 272.4, Urgent	0.69	nan	nan	29.42	nan	3.9	34.68	3.19

Figure 5.1: Resume table for the itemset analyzed in the population affected by *Depressive disorder*

= 311). For every itemset, the first column in Figure 5.1 shows the metalift and the remaining ones the *risk ratios* for the different factors analyzed, *'nan'* values represent cases in which the calculated *risk ratio* were not statistically significant. *Age* (≥ 40 and ≥ 65) and *Sex* (F) parameters were chosen by looking at the most representative of the population while the diagnostics by reviewing the literature looking for most common comorbidities of *depression*.

The chart showing the explanatory power of the different factors is also shown as output (Figure 5.2).

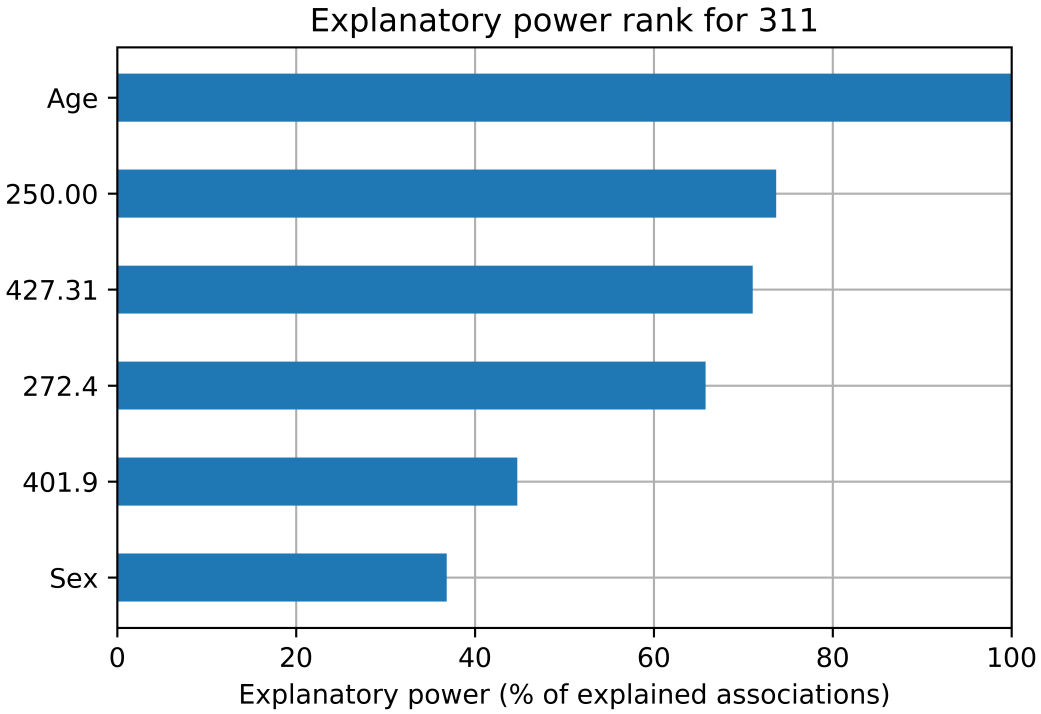


Figure 5.2: Explaining power of the analyzed factors in the population affected by *Depressive disorder*

Chapter 6

Results - Metalift

6.1 Metalift examples

As explained in section 5.1, $metalift$ value can range from 0 to infinite.

The cases in which $metalift < 1$ can be interpreted in different ways. For example the itemset

$$\{\textit{Minor diseases of respiratory system, Congestive heart failure}\}$$

showed a $lift_A = 3.76$ in the general population, but it decreased to $lift_B = 2.56$ when the population was restricted to people who suffers from *Hyperlipidemia* ($metalift = 0.68$). This information could be quite known, so not interesting, to healthcare specialists, since it is known that hyperlipidemia (i.e. high levels of cholesterol) in associated with COPD (Chronic Obstructive Pulmonary Disease) [39], that lead both to heart and respiratory diseases.

However, cases like the itemset

$$\{\textit{Minor diseases of respiratory system, Acute respiratory failure}\}$$

were observed. This itemset led to $metalift = 0.59$ when x=*Urinary tract infection (UTI)* was considered. The metalift value is similar to the first case, but we did not find much literature regarding a direct association between UTI and respiratory problems.

As a $metalift > 1$ example, the itemset

$$\{\textit{Knee joint replacement, Programmed visit to the hospital}\}$$

shows inverse association in base population, $lift_A = 0.84$, but turns out to be associated, $lift_B = 2.22$, in the population affected by Parkinson's disease, $metalift = 2.64$. Another example is the itemset

$$\{\textit{Long-term (current) use of insulin, Personal history of tobacco use}\}$$

that led to a $metalift = 1.60$ when it was considered in the population where x=*Pleural effusion*. The first case may be not interesting, because it could be explained by age, as we will show in section 5.3, but the second may be.

The cases where $metalift \approx 1$ are the ones that we want to filter, because they show no difference with respect to the whole population, and therefore tell nothing new about the subpopulation that the expert is examining today.

For example, itemset

$$\{\textit{Venous insufficiency, Atrial fibrillation}\}$$

showed $lift_A = 2.112$ in the general population and $lift_B = 2.118$ in the population where x = *acute posthemorrhagic anemia*, resulting in a $metalift = 1.003$.

Another example is the association between items

$$\{\textit{Old myocardial infarction, Chronic ischemic heart disease}\}$$

that are highly associated (quite obviously, since both are associated with heart problems) with a $lift_A = 12.55$ in the general population. This association persisted ($lift_B = 11.78$) also when it was analyzed in the population suffering from *Urinary tract infection*, confirming that this last disease is not particularly related to heart problems ($metalift = 1.06$).

These associations, as the cases where $lift \approx 1$ described in subsection 2.2.1, can be considered not interesting for a user that has chosen to study the population where the x is present. Therefore, these cases are some of the itemsets can be filtered from the analysis.

6.2 Filtering by metalift

The figures 6.1 and 6.2 show the results obtained by this method after being applied on the populations affected by *Hypertension* and x=*Tobacco use disorder* are showed, while the results obtained on the most frequent diagnostics of the dataset are listed in appendix C.

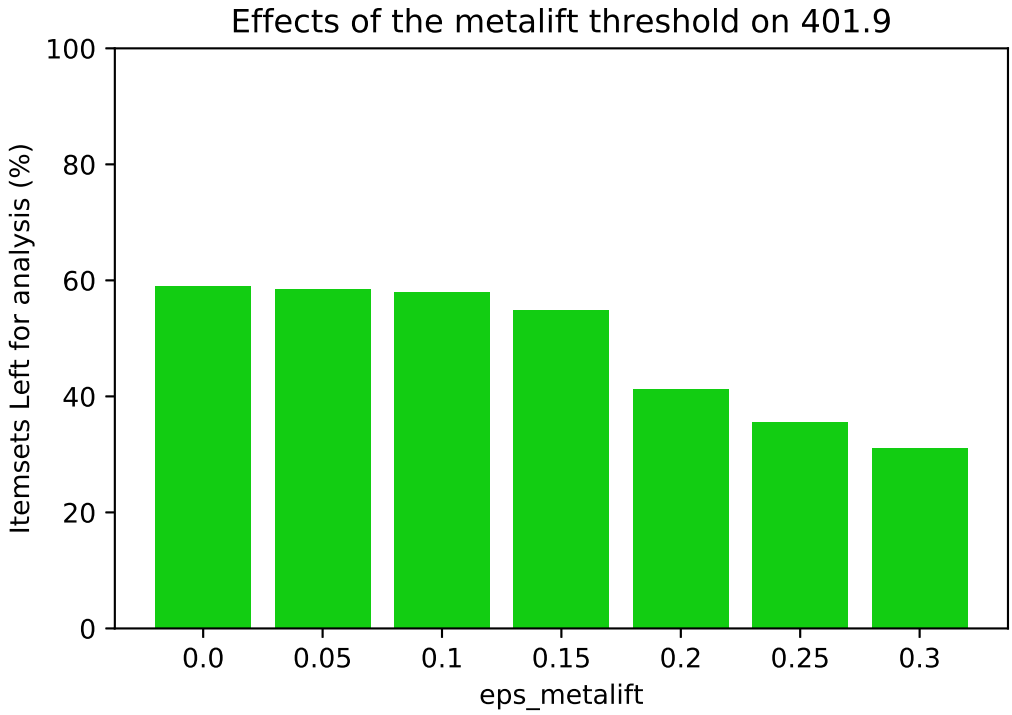


Figure 6.1: Effects of the metalift threshold on the population suffering from *Hypertension*

The first bar of each chart, the one showing the percentage of itemsets left after applying

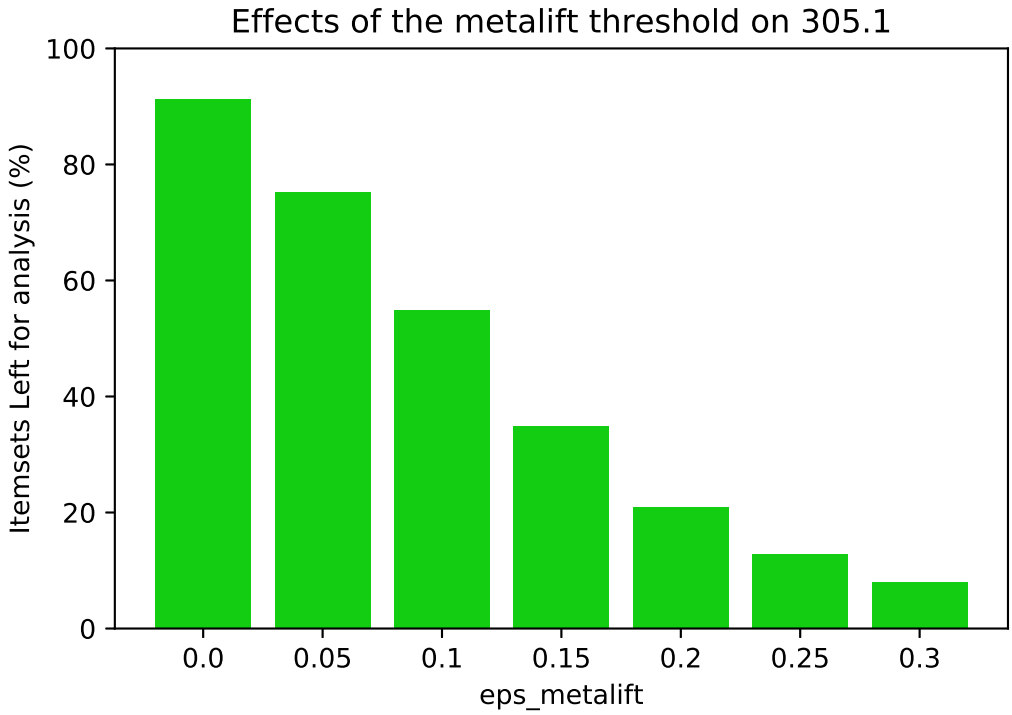


Figure 6.2: Effects of the metalift threshold on the population with *Tobacco use disorder*

$\epsilon_{metalift} = 0.0$, is not corresponding to the 100% as one might expect. This happens because, before applying the metalift filter to an x-affected population, we removed all the itemsets containing x since analyzing them would have had no reason, considering that it is obvious that an itemset containing x is strictly related to x.

From the two examples in the figures 6.1 and 6.2 it is possible to see that more itemset are filtered because of the presence of the x of interest in the case of *Hypertension* than in the case of *Tobacco use disorder*. That is explained by the fact that *401.9 - Hypertension* ($supp = 29.34\%$) is more frequent than *305.1 - Tobacco use disorder* ($supp = 9.54\%$) in the general population. The more acute effect of $\epsilon_{metalift}$ on *Tobacco use disorder* affected population than on *Hypertension* shows that *Hypertension* is mostly involved as a relevant factor in associations between diseases, leading to metalifts that are far from 1, while *Tobacco use disorder* leads more to $metalift \approx 1$ cases.

From the results above and from those listed in the appendix C, it is possible to understand that filtering by metalift has different effects depending on the populations a user is interested in, but this approach shows in general relevant reduction on the number of itemsets that has

to be analyzed.

Chapter 7

Results - Explanatory Variables

7.1 Explanatory variables

While looking for potential explanatory variables we analyzed the effects of several ones for each diagnostic x, by reviewing x-related literature or searching for most frequent variables in x-affected population. Age and sex were also analyzed for each case, and, as expected, they often showed high values for the *risk ratio* (see also values for age and sex attributes in Figure 5.1 in section 5.5), meaning that they are often important factors to explain an associations. However, we omitted the informations about them in the results explained in this sections (except for *Example 2*) because we preferred to focus on relationships between diseases and comorbidities.

Of course we obtained different effects and different risk ratio values depending on the factors we were analyzing, showing that some of them have a bigger impact on a given association than others. For simplicity, for each reported case, we report only the explanatory variable that turned out to have the biggest effects on the studied association. We show here the examples we consider to be more important or representative, other results are listed in appendix D.

Example 1: The itemset

$$I = \{Osteoporosis,\ Alzheimer's\ disease\}$$

showed a lift decrement by a $metalift = 0.57$ with x=*Diabetes mellitus*. Studies have actually reported that *Diabetes mellitus* is one of the risk factors associated to both *Osteoporosis* [40] and *Alzheimer's disease* [41], explaining the association in I.

Example 2: The itemset

$$\{Acquired\ absence\ of\ breast,\ Personal\ history\ of\ malignant\ neoplasm\ of\ breast\}$$

showed $metalift = 0.56$ when population suffering by *Depressive disorder* was analyzed. Knowing that *Depressive disorder* is more prevalent in women than in men [42], makes it easy to understand that the real factor (or one of the factors, the one that showed up by analyzing this population) that links those two diseases is being a female.

The prevalence of females over males in the *depressive disorder* affected population, or also the suggestion that sex could be one of the factors to analyze to understand an association, can be pointed out by doctors or healthcare specialists, but it can also be noticed by looking at the *support* of males and females in this population (population B). $Sex = F$ turns out to be prevalent over $Sex = M$ with a support ratio $\frac{supp(\{F\})}{supp(\{M\})} = 2.46$.

The explanatory role of $Sex = F$ in this example can be highlighted by calculating the Risk Ratio, setting $Sex = F$ as exposure and the said association as outcome. In this way the support of the I in the female population is compared to its support in the male population, confirming ($RR(Sex = F, I) > 1$ and with a significant *p-value*< 0.0001) that the association is partly explained by Sex.

Example 3: The case of

$$I = \{Pneumonitis\ due\ to\ inhalation,\ Acute\ respiratory\ failure\},$$

which showed $metalift = 1.34$ when the population affected by x =*Coronary atherosclerosis* was analyzed, but, looking at the Risk Ratios, no correlations between the items in I and x (or the factors related to x that we checked) justifying the metalift greater than 1 were found.

Example 4: The itemset

$$I = \{\textit{Glaucoma, Hyperlipidemia, Hypertension}\}$$

showed $metalift = 0.58$ for $x = \textit{Diabetes}$ and $RR(x, I) = 3.22$ with a $p\text{-}value < 0.0001$, showing that *diabetes* explains the association as confirmed by links between it and the three diagnostics in I [45, 46, 47].

The apparent conclusion is that if one reports the three itemsets

$$\{\textit{Glaucoma, Diabetes}\}, \quad \{\textit{Hyperlipidemia, Diabetes}\}, \quad \textit{and} \quad \{\textit{Hypertension, Diabetes}\}$$

one need not report the itemset {Glaucoma, Hyperlipidemia, Hypertension}.

The catch here is that we interpret this finding in the sense that *Diabetes* is the explanatory variable, and the other three he explained ones. But it could also happen that, similarly, *Glaucoma* numerically explains the association between *Diabetes*, *Hypertension*, and *Hyperlipidemia*, or in general, that the direction of *"explanation"* could go in several ways. Yet, every single clinician will confirm that it is *Diabetes* that explains anything related to *Glaucoma*. You do not get *Diabetes because* you have *Glaucoma*, it goes the other way round.

Additional information such as temporal information, causal information, or a notion of *"seriousness"* of a disease provided by clinicians could be used to break symmetries in an automated way. This is a very interesting avenue for future work. Still, see the next section on Explanatory Power for a first approach to determine what the most central, serious, or explanatory variables are.

Example 5: In the population affected by $x = \textit{Hypertensive chronic kidney disease}$, the itemset

$$I = \{\textit{ Use of anticoagulants, Acute respiratory failure, Hyperlipidemia, Urgent admission}\}$$

showed a $metalift = 0.53$ and $y =$Atrial fibrillation was found to be strictly linked to x [48] and showed $RR(y, I) = 34.68$ and a 95% CI=[25.63, 46.92]. A review of the literature confirmed

the relationships between *Atrial fibrillation* and the diagnostics in I ([49, 50, 51]) and that it could explain I

Example 6: The itemset

$$I = \{\ \textit{Hearing loss, Hyperlipidemia, Urgent admission}\}$$

showed a $metalift = 0.55$ when the population affected by $x =$*Hypertensive chronic kidney disease* was analyzed.

The association between diagnostics in I is still unclear [60].

With our approach we found a possible link between the two diseases, since $d =$*Chronic kidney disease*, that is the general case of x, has a $RR(d, I) = 3.07$ with a 95% confidence interval of [2.23, 4.23] and could explain this association. Literature confirming the relationship between both d and *Hearing loss* [58] or d and *Hyperlipidemia* [59] was found.

In general, strong relationships between cardiovasculatory, respiratory and kidney problems were the main ones that often appeared through the analysis. These are well known associations among healthcare specialists [75], and keeping those into account we were able to reveal their explanatory role in most of the itemsets we analyzed. In particular our approach was able to reveal the explanatory factors also when analyzing longer itemsets (containing 3 or more diagnostics) as stated in examples 4, 5, 6.

7.2 Explanatory Power

As described in section 5.4, we computed a rank of the most explanatory factors and comorbidities for a given diagnostic x by looking at the amount of important associations for x that a factor/association explains.

To obtain these results, we did not keep track of the values obtained from the metalift and the risk ratio of each factor involved in the analysis (as we did in section 7.1, where we showed

only the most important ones), so that a factor is either classified as explanatory or not for an association, without quantitative informations.

So, to clarify, a factor f has *explanatory power* $= 10$ for diagnostic x if there are 10 important associations among the important ones for x that are explained by f.

Some of the main results are showed in this section, other findings are listed in appendix E.

In figure 7.1 the rank for *Depression* affected population is showed. As expected age is the factor that explains most of the associations studied for the population. In particular, we point out that *Diabetes* and *Atrial fibrillation* turn out to be two of the most important comorbidities, explaining about the 70% of the associations. This is a result that was found in the literature, that suggests how, *"due to potential negative health consequences associated with comorbid diabetes and depression, both conditions should be optimally treated to maximize patient outcome"* [71] and that *"to achieve more comprehensive atrial fibrillation (AF) symptom relief, treatment of both AF and psychological comorbidities may be beneficial."* [70].

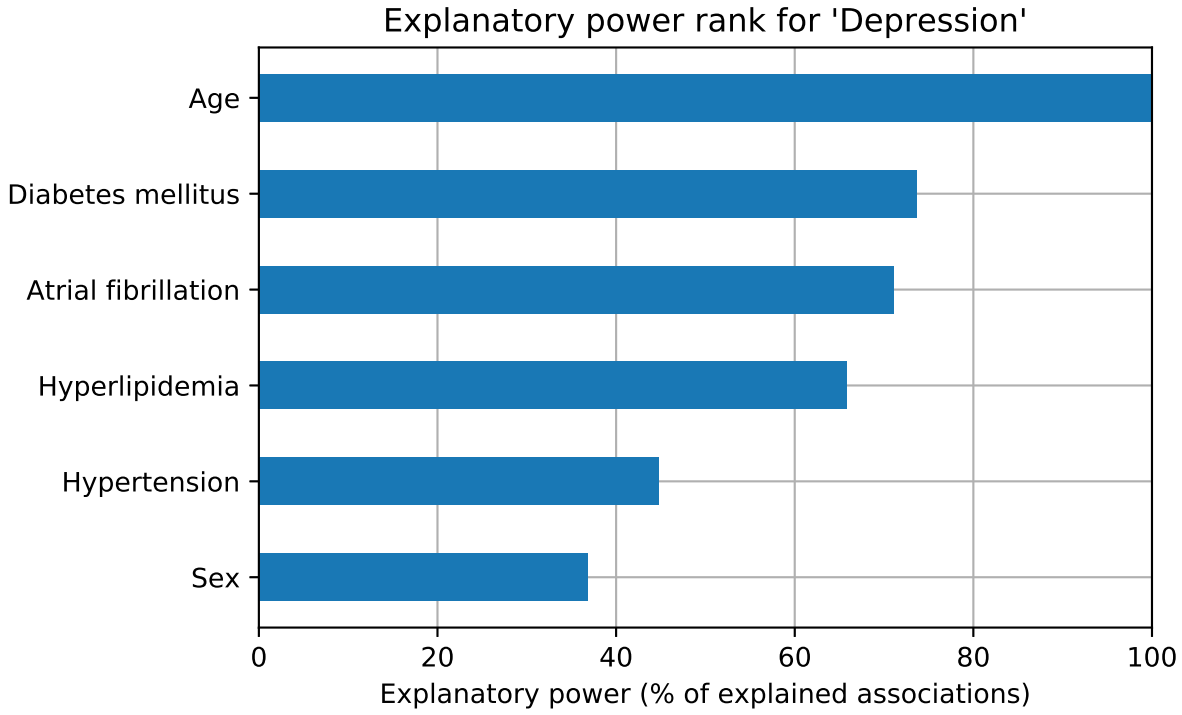


Figure 7.1: Explanatory power of different factors for depression-affected population

The figure 7.2 shows again the close dependency of *diabetes* by age, that is involved in all the association analyzed for this population.

Our approach was able to reveal the strongest comorbidities of *diabetes* [72], like *hypertension, hyperlipidemia* or *atrial fibrillation*. We were not able to find a relevant explanatory power for sex in this case but [73] states that *"It is often assumed that there is little or no sex bias within either Type I (insulin-dependent) or Type II (non-insulin-dependent) diabetes mellitus"*, so this is an information that may need to be checked by specialists.

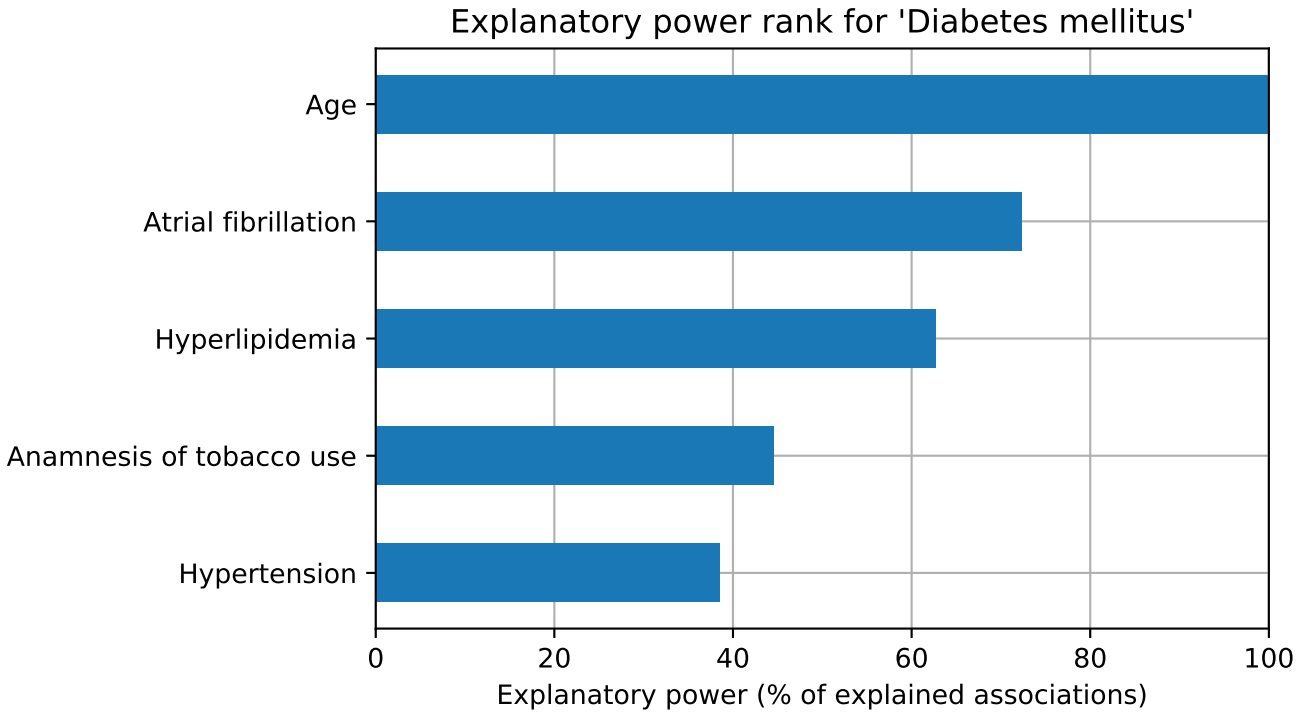


Figure 7.2: Explanatory power of different factors for diabetic population

From the rank for the explanatory power of factors related to *Urinary trait infection* (UTI) in figure 7.3 it is possible to see that, apart from age explaining almost all the associations found, *Escherichia Coli* is a factor that significantly affects more that 30% of the relationships, this because, as stated by [74] *"Escherichia coli is the most predominant pathogen causing 80-90% of community-acquired UTIs"*.

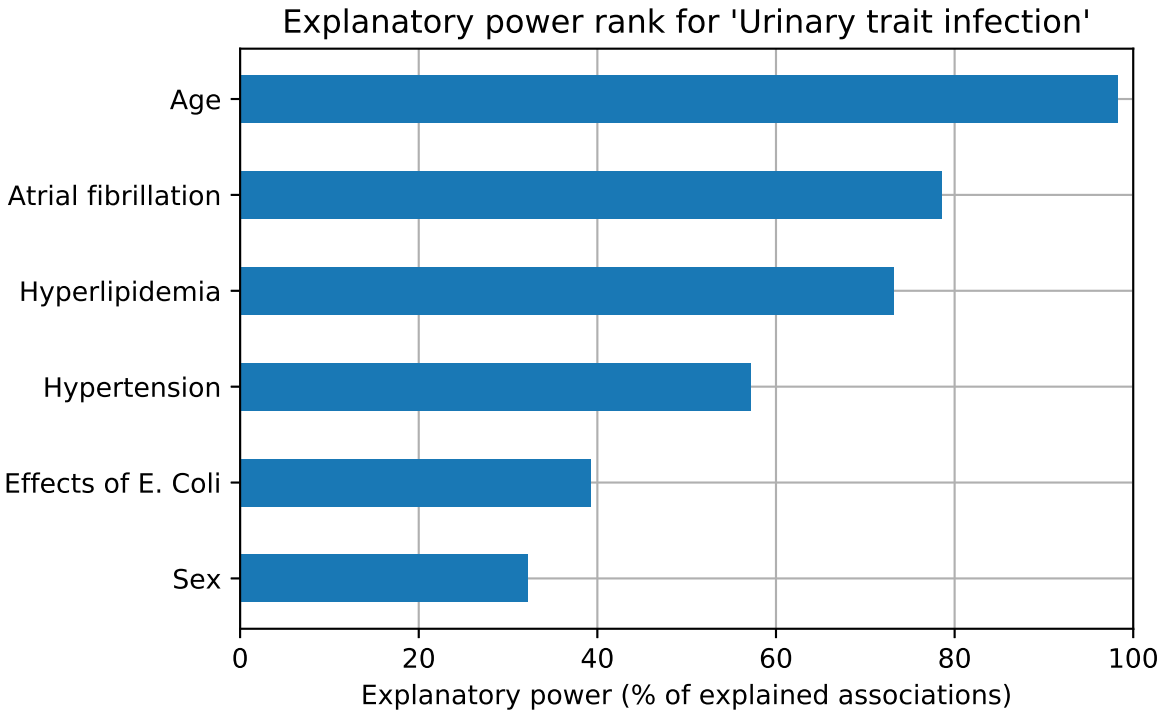


Figure 7.3: Explanatory power of different factors for population suffering of urinary trait infection

Chapter 8

Conclusion and Future Work

8.1 Summary of Thesis Achievements

The application of the *metalift*-based filtering method showed that interesting results can be obtained when looking for associations that have nothing to do with a disease of interest.

After filtering the not-interesting rules for several populations of interest our approach was able to filter, on average, 28.08% of the mined itemsets, with the best case of 69.55% associations filtered for the population affected by *Tobacco use disorder*. This, combined to other filtering methods, can help to reduce the huge amount of itemsets that are commonly produced as an output of a pattern mining algorithm, helping domain (in this case healthcare) specialists in the analysis.

By means of the *explanatory variable* study, instead, we were able to reveal more potentially not-interesting associations, also providing statistical significance informations about the variables explaining associations. By ranking the different factors we analyzed for each population of interest we managed to point out which are the most important comorbidities for a disease. These informations can help doctors and specialists in the decision-making process, giving suggestions on which comorbidities a greater attention should be given to while taking care of a patient

However, all the *explanatory power* rankings we found some results that are already reported in the general healthcare literature. This is probably due to the fact that we based our choice for factors and comorbidities to analyze only on reviews of the existent literature, finding only obvious or well known links between diagnostics.

As already explained in this report, help of healthcare specialists is required to improve the results, since, in this case of healthcare analysis but also in general, the notion of interestingness is mostly subjective and strictly linked to domain-specific knowledge.

8.2 Future Work

The main future step of this project concerns the development of a software to automatically analyze the given data from the preprocessing step and to the display of the results.

Given a graphic user interface for the software we developed, a given specialist should be able to select a population of interest and tune the parameters used to modify the filters applied to the mined associations.

The list of explanatory factors to analyze could be chosen by the user, but we also intend to automate the search for explanatory variables, showing the user only the real and most important ones.

Bibliography

[1] Agrawal R, Imielinski T, Swami A (1993) *Mining association rules between sets of items in large databases.* In: Proceedings of the 1993 ACM-SIGMOD international conference on management of data (SIGMOD93), Washington, DC, pp 207216

[2] Rakesh Agrawal , Ramakrishnan Srikant, *Fast Algorithms for Mining Association Rules in Large Databases* , Proceedings of the 20th International Conference on Very Large Data Bases, p.487-499, September 12-15, 1994

[3] S. Rao, R. Gupta, *Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm*, International Journal of Computer Science And Technology, pp. 489-493, Mar. 2012

[4] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining* Knowledge and Information Systems, vol. 14, no. 1, pp. 137, Dec. 2007.

[5] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. 1995. *An effective hash-based algorithm for mining association rules.* In Proceedings of the 1995 ACM SIGMOD international conference on Management of data (SIGMOD '95), Michael Carey and Donovan Schneider (Eds.). ACM, New York, NY, USA, 175-186

[6] Hannu Toivonen. 1996. *Sampling Large Databases for Association Rules.* In Proceedings of the 22th International Conference on Very Large Data Bases (VLDB '96), T. M. Vijayara-

man, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 134-145.

[7] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. 1997. *Dynamic itemset counting and implication rules for market basket data.* SIGMOD Rec. 26, 2 (June 1997), 255-264.

[8] Cheung DW, Han J, Ng VT, Wong CY (1996) *Maintenance of discovered association rules in large databases: an incremental updating approach.* In: Proceedings of the twelfth international conference, data engineering, pp 106114

[9] Rakesh Agrawal , John C. Shafer, *Parallel Mining of Association Rules*, IEEE Transactions on Knowledge and Data Engineering, v.8 n.6, p.962-969, December 1996

[10] Mohammed J. Zaki , Srinivasan Parthasarathy , Mitsunori Ogihara , Wei Li. *Parallel Algorithms for Discovery of Association Rules*, Data Mining and Knowledge Discovery, v.1 n.4, p.343-373, December 1997

[11] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. *Mining frequent patterns without candidate generation.* In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). ACM, New York, NY, USA, 1-12.

[12] M.S. Mythili and A.R. Mohamed Shanavas, *Performance Evaluation of Apriori and FP-Growth Algorithms* in International Journal of Computer Applications (0975 8887) Volume 79 No10, October 2013

[13] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. 2000. *Depth first generation of long patterns.* In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00). ACM, New York, NY, USA, 108-118.

[14] Jian Pei, Jiawei Han, Hongjun Lu, Shojiro Nishio, Shiwei Tang, Dongqing Yang. (n.d.). *H-mine: hyper-structure mining of frequent patterns in large databases.* In Proceedings 2001 IEEE International Conference on Data Mining. IEEE Comput. Soc

[15] Junqiang Liu, Yunhe Pan, Ke Wang, and Jiawei Han. 2002. *Mining frequent item sets by opportunistic projection.* In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 229-238.

[16] Zaki, M. J. (2000). *Scalable algorithms for association mining.* IEEE Transactions on Knowledge and Data Engineering, 12(3), 372390.

[17] Holsheimer M, Kersten M, Mannila H, Toivonen H (1995): *A perspective on databases and data mining.* In Proceeding of the 1995 international conference on knowledge discovery and data mining (KDD95), Montreal, Canada, pp 150155

[18] Pasquier N., Bastide Y., Taouil R., Lakhal L. (1999) *Discovering Frequent Closed Itemsets for Association Rules.* In: Beeri C., Buneman P. (eds) Database Theory ICDT99. ICDT 1999. Lecture Notes in Computer Science, vol 1540. Springer, Berlin, Heidelberg

[19] Roberto J. Bayardo, Jr.. 1998. *Efficiently mining long patterns from databases.* In Proceedings of the 1998 ACM SIGMOD international conference on Management of data (SIGMOD '98), Ashutosh Tiwary and Michael Franklin (Eds.). ACM, New York, NY, USA, 85-93.

[20] Pei, J., Han, J., and Mao, R. 2000.*CLOSET: An efficient algorithm for mining frequent closed itemsets.* In Proc. 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery (DMKD'00), Dallas, TX, pp. 1120.

[21] Grahne, Gsta, Zhu, Jianfei. (2003). *Efficiently Using Prefix-trees in Mining Frequent Itemsets.* FIMI'03 Workshop on Frequent Itemset Mining Implementations: 2003.

[22] Wang, J., Han, J., Pei, J. (2003). *CLOSET+: Searching for the best strategies for mining frequent closed itemsets.* In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03 (pp. 236-245).

[23] Zaki, M. J., Hsiao, C.-J. (2002). *CHARM: An Efficient Algorithm for Closed Itemset Mining.* In Proceedings of the 2002 SIAM International Conference on Data Mining (pp. 457473). Society for Industrial and Applied Mathematics.

[24] Ken McGarry. 2005. *A survey of interestingness measures for knowledge discovery.* Knowl. Eng. Rev. 20, 1 (March 2005), 39-61.

[25] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. 2002. *Selecting the right interestingness measure for association patterns.* In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 32-41.

[26] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S. (2013, March). *Using of Jaccard coefficient for keywords similarity.* In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, No. 6).

[27] Nguyen, H. V., Bai, L. (2010, November). *Cosine similarity metric learning for face verification.* In Asian conference on computer vision (pp. 709-720). Springer, Berlin, Heidelberg.

[28] Sigal Sahar. 1999. *Interestingness via what is not interesting.* In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '99). ACM, New York, NY, USA, 332-336.

[29] Avi Silberschatz and Alexander Tuzhilin. 1995. *On subjective measures of interestingness in knowledge discovery.* In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD'95), Usama Fayyad and Ramasamy Uthurusamy (Eds.). AAAI Press 275-281.

[30] Koh HC, Tan G. *Data mining applications in healthcare.* J Healthc Inf Manag. 2005;19(2):64-72.

[31] Boytcheva S., Angelova G., Tcharaktchiev D., Angelov Z. *Big Data Analytics in Healthcare Pattern Mining of Temporal Clinical Events*, 2001.

[32] Wright AP, Wright AT, Mccoy AB, Sittig DF. *The use of sequential pattern mining to predict next prescribed medications.* J Biomed Inform. 2015;53:73-80.

[33] Boytcheva, S. et al. (2017). *Mining comorbidity patterns using retrospective analysis of big collection of outpatient records.* Health Information Science and Systems, 5(1).

[34] Kim, H.S. et al. (2012). *Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining.* The Korean Journal of Internal Medicine, 27(2), 197.

[35] SearchHealtIT - *ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification).* Retrieved from https://searchhealthit.techtarget.com/definition/ICD-10-CM

[36] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, and Milos Hauskrecht. 2013. *A temporal pattern mining approach for classifying electronic health record data.* ACM Trans. Intell. Syst. Technol. 4, 4, Article 63 (October 2013), 22 pages.

[37] Chen, Y., Xu, R. (2014). *Mining Cancer-Specific Disease Comorbidities from a Large Observational Health Database.* Cancer Informatics, 13(Suppl 1), 3744.

[38] Borgelt, C. (n.d.). *Eclat.* Retrieved from http://www.borgelt.net/doc/eclat/eclat.html

[39] Reed, Robert M. et al. *Advanced chronic obstructive pulmonary disease is associated with high levels of high-density lipoprotein cholesterol.* The Journal of Heart and Lung Transplantation , Volume 30 , Issue 6 , 674 - 678

[40] Stavroula Paschou, Anastasia D Dede, Panagiotis G Anagnostis, Andromachi Vryonidou, Daniel Morganstein, Dimitrios G Goulis; *Type 2 Diabetes and Osteoporosis: A Guide to Optimal Management* , The Journal of Clinical Endocrinology and Metabolism, Volume 102, Issue 10, 1 October 2017, Pages 36213634

[41] Arvanitakis Z, Wilson RS, Bienias JL, Evans DA, Bennett DA. *Diabetes Mellitus and Risk of Alzheimer Disease and Decline in Cognitive Function.* Arch Neurol. 2004;61(5):661666.

[42] Albert, P. R. (2015). *Why is depression more prevalent in women?* Journal of Psychiatry and Neuroscience: JPN, 40(4), 219221. http://doi.org/10.1503/jpn.150205

[43] Ekechi Lucky Chinedu. *Itemset mining in electronic healthcare records: pruning redundancy by considering explanatory variables.* Master thesis, program on Artificial Intelligence, Department of Computer Science, Universitat Politcnica de Catalunya, 2017.

[44] Morris, J. A., Gardner, M. J. (1988). *Statistics in Medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates.* British Medical Journal (Clinical Research Ed.), 296(6632), 13131316.

[45] Zhao, Y.-X., Chen, X.-W. (2017). *Diabetes and risk of glaucoma: systematic review and a Meta-analysis of prospective cohort studies.* International Journal of Ophthalmology, 10(9), 14301435.

[46] OBrien, T., Nguyen, T. T., Zimmerman, B. R. (1998). *Hyperlipidemia and Diabetes Mellitus.* Mayo Clinic Proceedings, 73(10), 969976.

[47] Schutta MH. *Diabetes and hypertension: epidemiology of the relationship and pathophysiology of factors associated with these comorbid conditions.* J Cardiometab Syndr. 2007;2(2):124-30.

[48] Tsiachris D, Tsioufis C, Mazzone P, Katsiki N, Stefanadis C. *Atrial fibrillation and chronic kidney disease in hypertension: a common and dangerous triad.* Curr Vasc Pharmacol. 2015;13(1):111-20.

[49] Katritsis DG, Gersh BJ, Camm AJ. *Anticoagulation in Atrial Fibrillation Current Concepts.* Arrhythmia and Electrophysiology Review. 2015;4(2):100-107.

[50] Hightower O. Atrial Fibrillation and Acute Respiratory Failure: Unique Presentation of Diffuse Large B-Cell Lymphoma. The Ochsner Journal. 2014;14(2):248-251.

[51] Zheng J, Shi L, Xu W, et al. *Impact of hyperlipidemia and atrial fibrillation on the efficacy of endovascular treatment for acute ischemic stroke: a meta-analysis.* Oncotarget. 2017;8(42):72972-72984

[52] Johnson JR. *Virulence factors in Escherichia coli urinary tract infection.* Clin Microbiol Rev. 1991;4(1):80-128.

[53] Hall SA, Chiu GR, Kaufman DW, Wittert GA, Link CL, McKinlay JB. *Commonly-used antihypertensives and lower urinary tract symptoms: Results from the Boston Area Community Health (BACH) Survey.* Bju International. 2012;109(11):1676-1684

[54] Alonso A, Arenas de Larriva AP. *Atrial Fibrillation, Cognitive Decline And Dementia.* European cardiology. 2016;11(1):49-53. doi:10.15420/ecr.2016:13:2.

[55] Gill S, Jun M, Ravani P. *Atrial fibrillation and chronic kidney disease: struggling through thick and thin.* Nephrol Dial Transplant. 2017;32(7):1079-1084.

[56] Lee Park K, Anter E. *Atrial Fibrillation and Heart Failure: A Review of the Intersection of Two Cardiac Epidemics.* Journal of Atrial Fibrillation. 2013;6(1):751.

[57] Einhorn LM, Zhan M, Hsu VD, et al. *The frequency of hyperkalemia and its significance in chronic kidney disease.* Archives of internal medicine. 2009;169(12):1156-1162. doi:10.1001/archinternmed.2009.132.

[58] *The Association Between Reduced GFR and Hearing Loss: A Cross-sectional Population-Based Study* Vilayur, Eswari et al. American Journal of Kidney Diseases , Volume 56 , Issue 4 , 661 - 669

[59] Sahadevan M, Kasiske BL. Hyperlipidemia in kidney disease: causes and consequences. Curr Opin Nephrol Hypertens. 2002;11(3):323-9.

[60] Evans MB, Tonini R, Shope CD, et al. *Dyslipidemia and Auditory Function.* Otology and neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology. 2006;27(5):609-614.

[61] Courties A, Sellam J. *Osteoarthritis and type 2 diabetes mellitus: What are the links?.* Diabetes Res Clin Pract. 2016;122:198-206.

[62] Sun Y, Hu D. *The link between diabetes and atrial fibrillation: cause or correlation?* Journal of Cardiovascular Disease Research. 2010;1(1):10-11.

[63] Caples SM, Somers VK. *Sleep Disordered Breathing and Atrial Fibrillation.* Progress in cardiovascular diseases. 2009;51(5):411-415.

[64] Nelson RH. *Hyperlipidemia as a Risk Factor for Cardiovascular Disease.* Primary care. 2013;40(1):195-211.

[65] Yu S, Christiani DC, Thompson BT, Bajwa EK, Gong MN. *Role of Diabetes in the Development of Acute Respiratory Distress Syndrome.* Critical care medicine. 2013;41(12):2720-2732.

[66] Bautista LE, Vera LM. *Antihypertensive effects of aspirin: what is the evidence?.* Curr Hypertens Rep. 2010;12(4):282-9.

[67] Ames RP. *Hyperlipidemia in hypertension: causes and prevention.* Am Heart J. 1991;122(4 Pt 2):1219-24.

[68] Schnabel RB, Michal M, Wilde S, et al. *Depression in Atrial Fibrillation in the General Population.* Watz H, ed. PLoS ONE. 2013;8(12):e79109.

[69] Lau C-P. *Pacing for atrial fibrillation.* Heart. 2003;89(1):106-112.

[70] Thompson TS, Barksdale DJ, Sears SF, Mounsey JP, Pursell I, Gehi AK. *The effect of anxiety and depression on symptoms attributed to atrial fibrillation.* Pacing Clin Electrophysiol. 2014;37(4):439-46.

[71] Katon WJ. *The Comorbidity of Diabetes Mellitus and Depression.* The American journal of medicine. 2008;121(11 Suppl 2):S8-15.

[72] Jelinek HF, Osman WM, Khandoker AH, et al. *Clinical profiles, comorbidities and complications of type 2 diabetes mellitus in patients from United Arab Emirates.* BMJ Open Diabetes Research and Care. 2017;5(1):e000427.

[73] Gale EA, Gillespie KM. *Diabetes and gender.* Diabetologia. 2001;44(1):3-15.

[74] Ejrnaes K. *Bacterial characteristics of importance for recurrent urinary tract infections caused by Escherichia coli.* Dan Med Bull. 2011;58(4):B4187.

[75] Faeq Husain-Syed, Peter A. McCullough, Horst-Walter Birk, Matthias Renker, Alessandra Brocca, Werner Seeger, Claudio Ronco, *Cardio-Pulmonary-Renal Interactions: A Multidisciplinary Approach*,Journal of the American College of Cardiology,Volume 65, Issue 22,2015, Pages 2433-2448,

Appendices

Appendix A

ICD9 codes

The ICD9 codes used in this report and their description are reported in the table below.

ICD9 code	**Description**
041.49	Effects of E. Coli
244.9	Hypothyroidism
250.00	Diabetes mellitus
272.4	Hyperlipidemia
276.7	Hyperpotassemia
278.00	Obesity
294.10	Dementia
300.4	Dysthymic disorder
305.1	Tobacco use disorder
311	Depression
365.9	Glaucoma
389.9	Hearing loss
401.9	Hypertension
403.90	Hypertensive chronic kidney disease
414.01	Coronary atherosclerosis

414.9	Chronic ischemic heart disease
426.3	Left bundle branch block
427.31	Atrial fibrillation
428.0	Heart failure
496	Chronic airway obstruction
518.81	Respiratory failure
518.84	Acute and chronic respiratory failure
553.3	Diaphragmatic hernia
584.9	Acute kidney failure
585.3	Chronic kidney disease (moderate)
585.9	Chronic kidney disease
599.0	Urinary trait infection
715.90	Osteoarthrosis
733.00	Osteoporosis
V10.3	Personal history of malignant neoplasm of breast
V15.82	Anamnesis of tobacco use
V45.01	Cardiac pacemaker
V45.71	Acquired absence of breast and nipple
V58.61	Long term use of anticoagulants
V58.66	Long term use of aspirin
V58.67	Long term use of insulin

Table A.1: ICD9 codes and their respective description

Appendix B

Diagnostics prevalences

The most prevalent diseases for the studied population when stratified by age and sex are listed below.

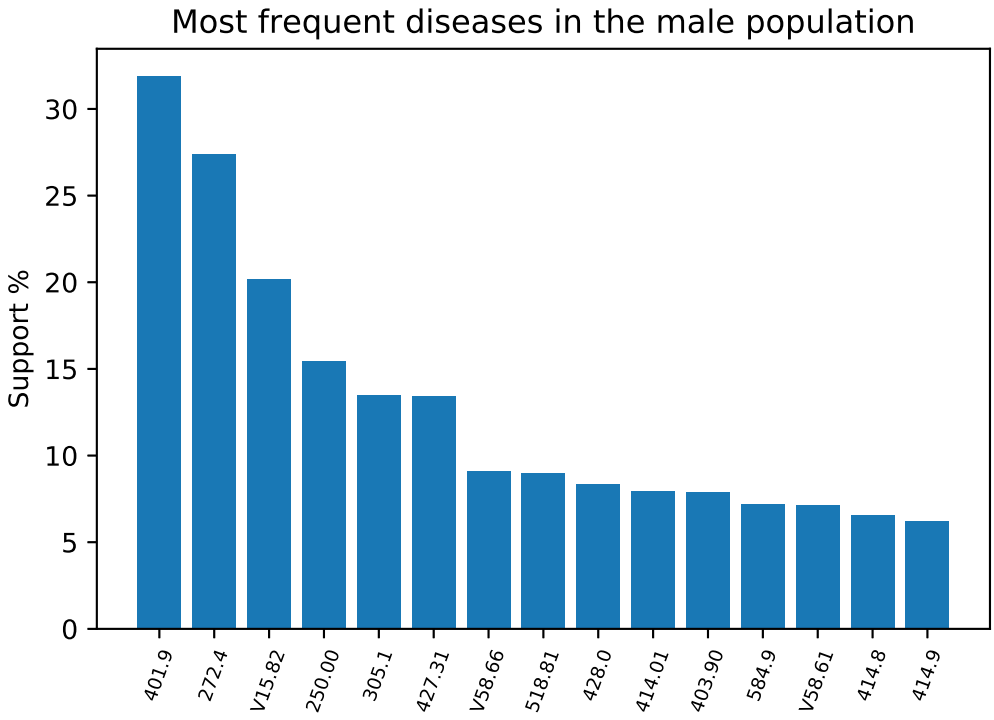


Figure B.1: Diagnostics in the male population

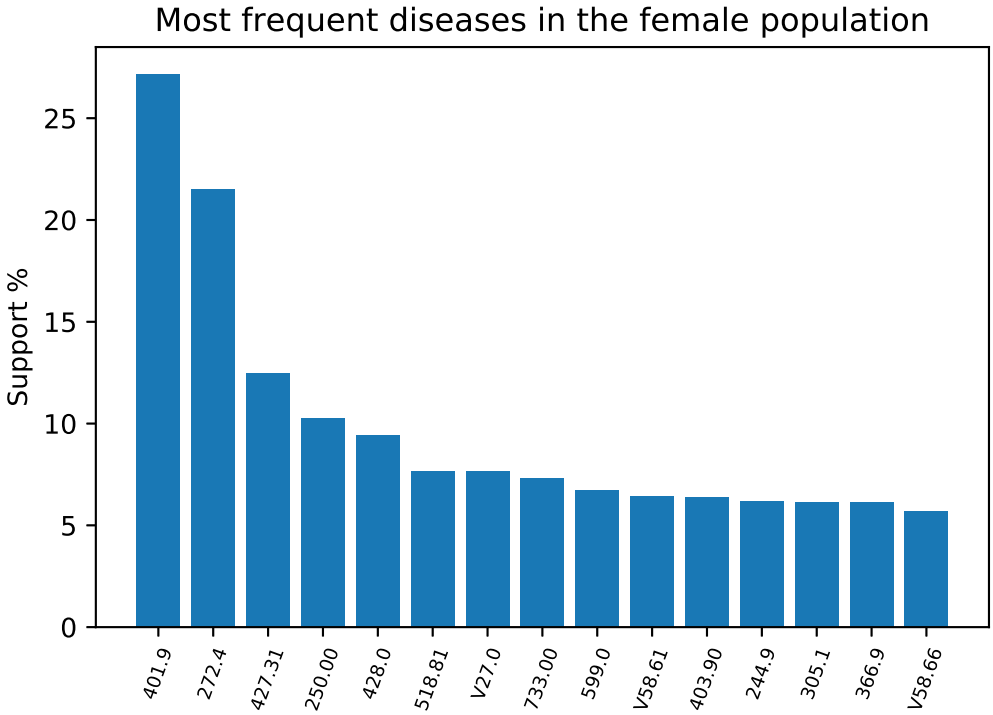


Figure B.2: Diagnostics in the female population

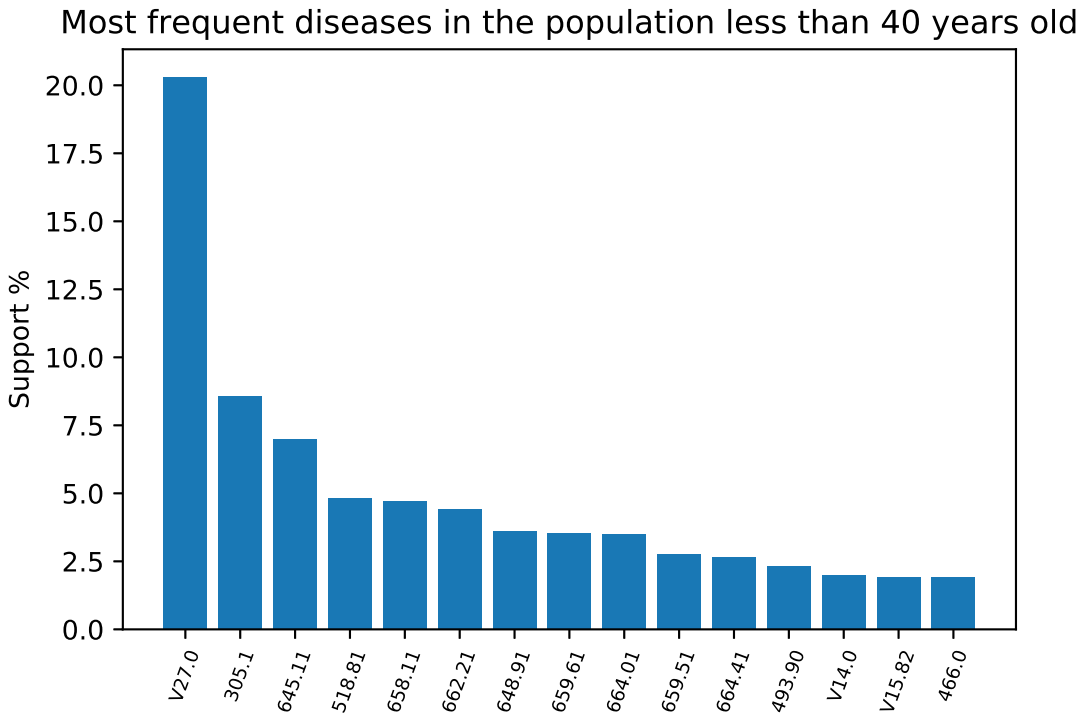


Figure B.3: Diagnostics in the population below 40 years old

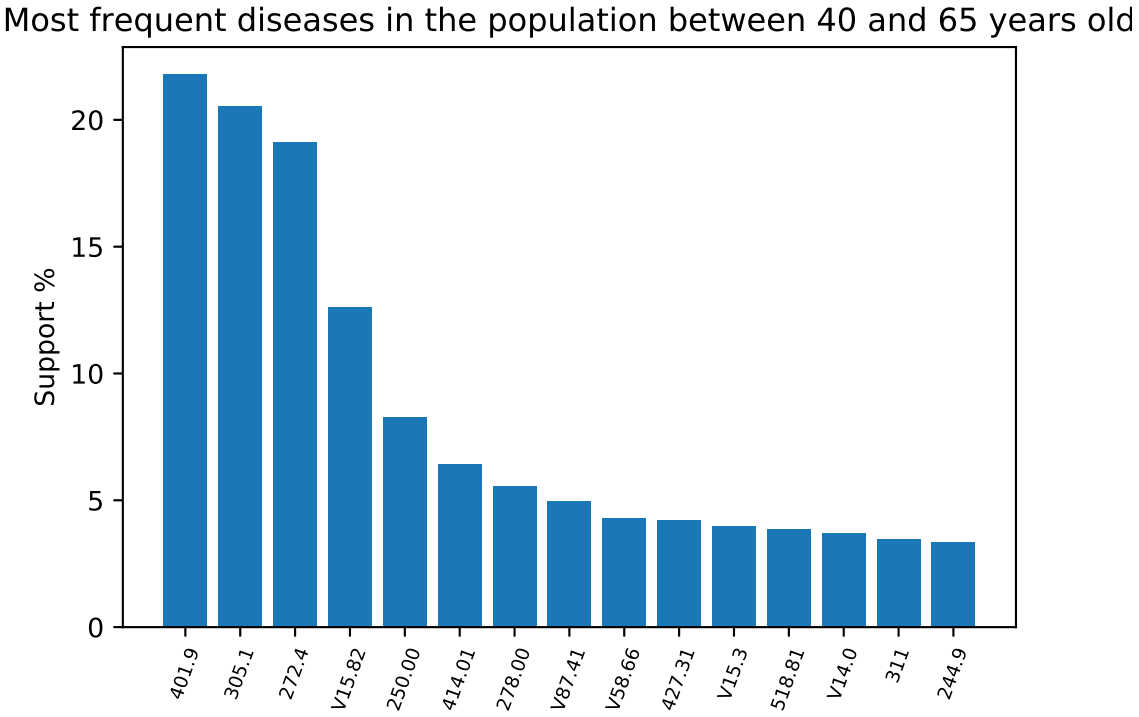


Figure B.4: Diagnostics in the population between 40 and 65 years old

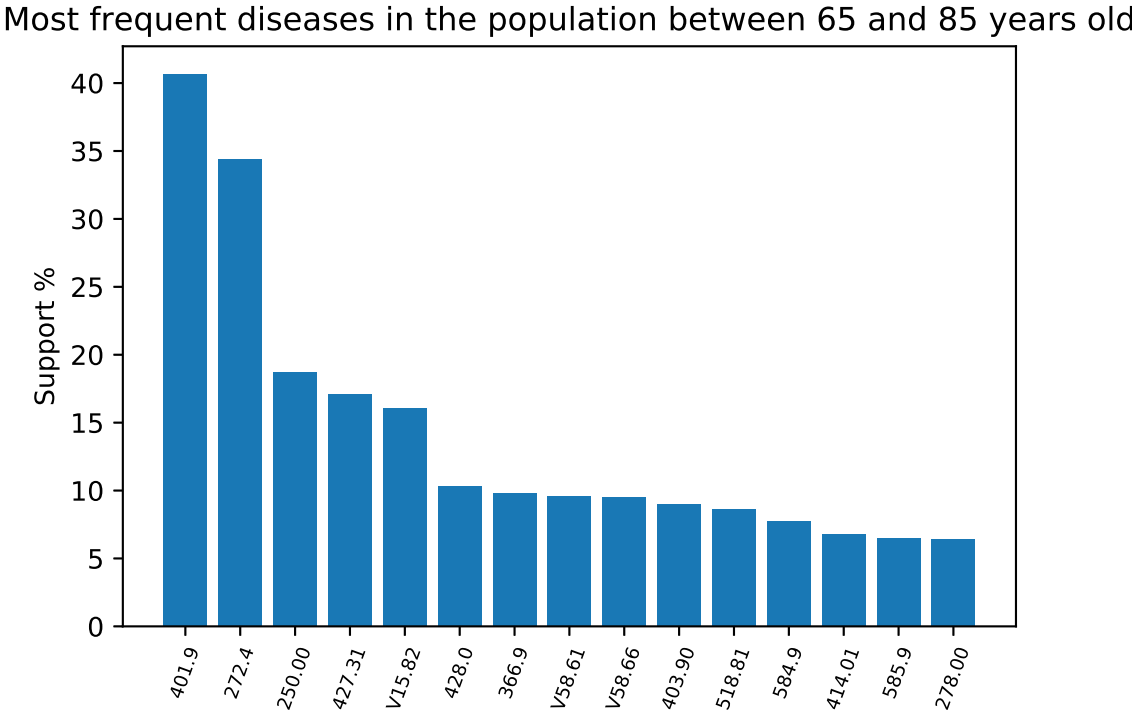


Figure B.5: Diagnostics in the population between 65 and 85 years old

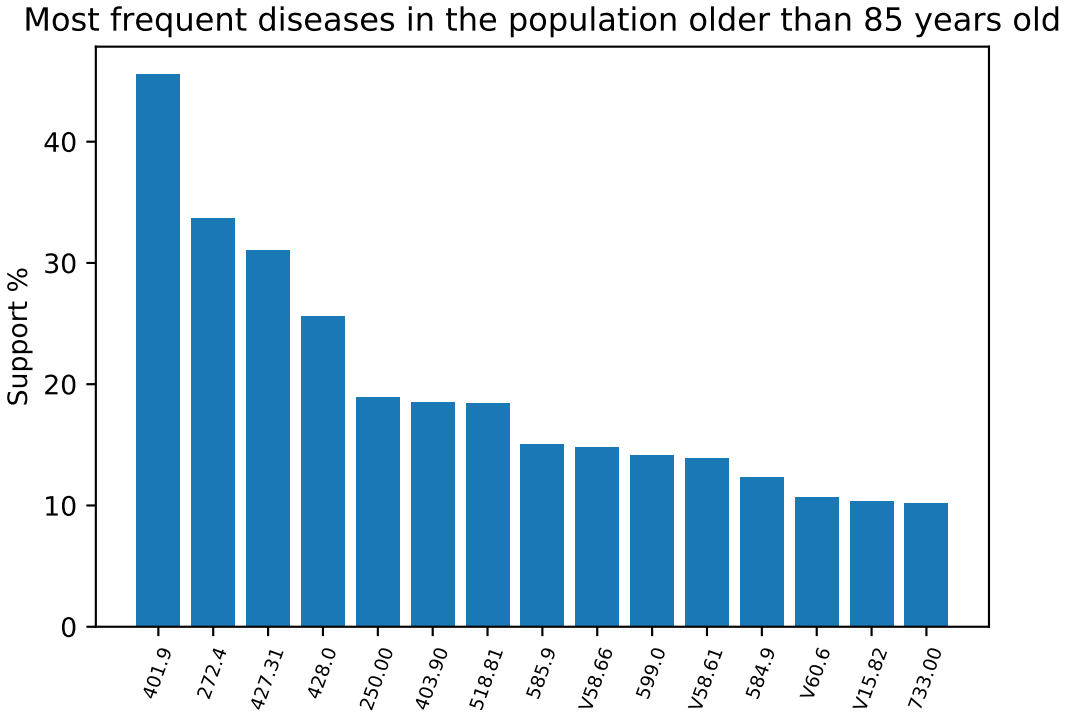
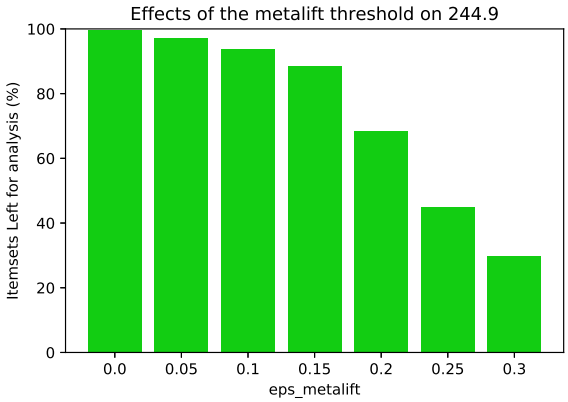


Figure B.6: Diagnostics in the population above 85 years old

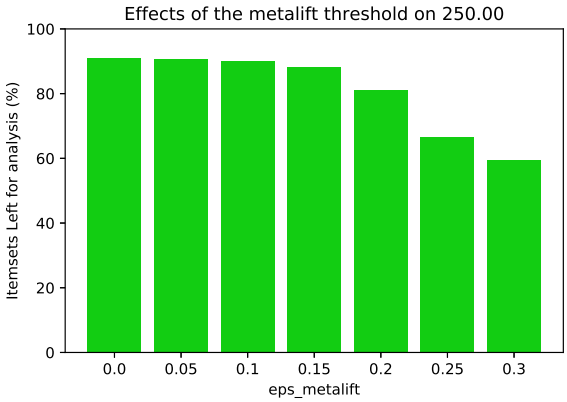
Appendix C

Effects of filtering by metalift on different diagnostics

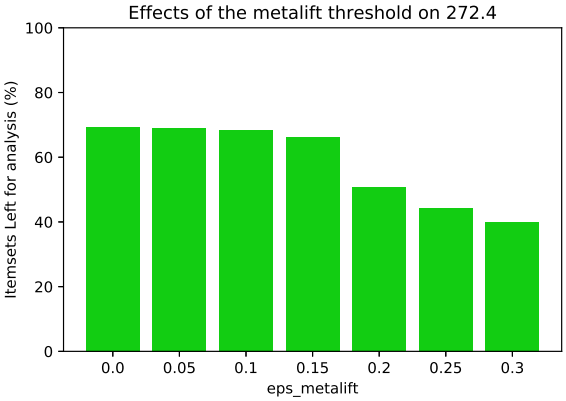
The effects of $\epsilon_{metalift}$ on the number of itemsets to analyze after filtering by metalift was applied to different x-affected populations are shown in the charts below. As expected, the cases in which the percentage of itemsets removed before applying the filter $\epsilon_{metalift} = 0.0$, with the criteron explained in section 5.2 is higher are the ones where x corresponds to the most frequent diagnostics as *Hyperlipidemia* (272.4) or *Hypertension* (401.9).



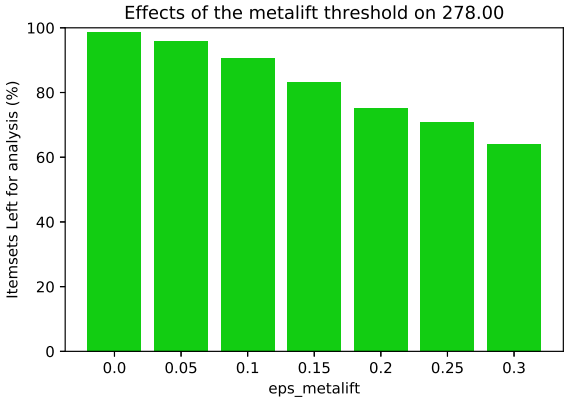
(a) 244.9



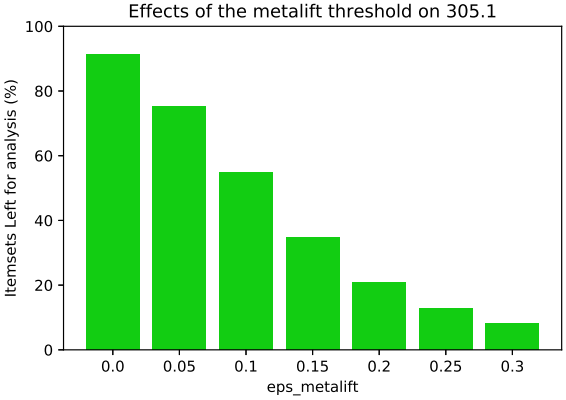
(b) 250.00



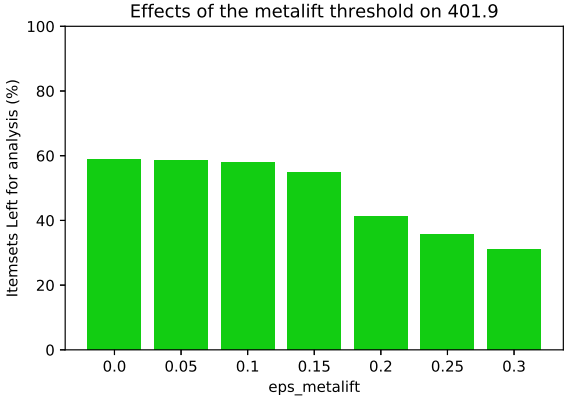
(c) 272.4



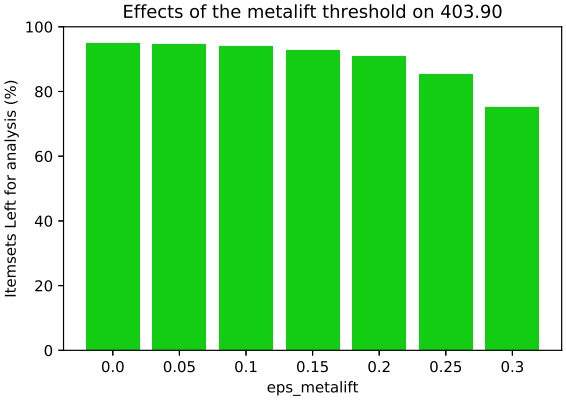
(d) 278.00



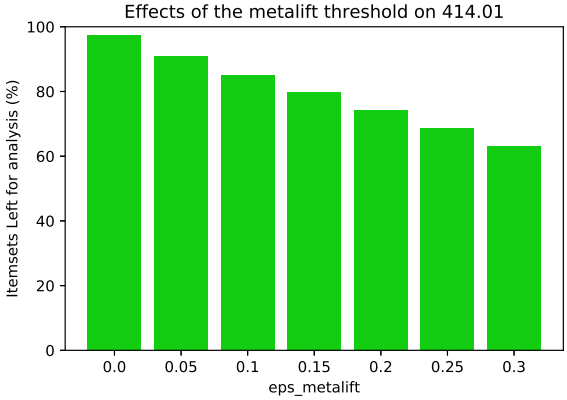
(e) 305.1



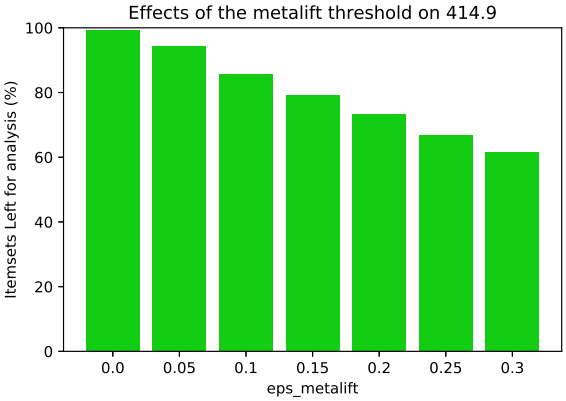
(f) 401.9



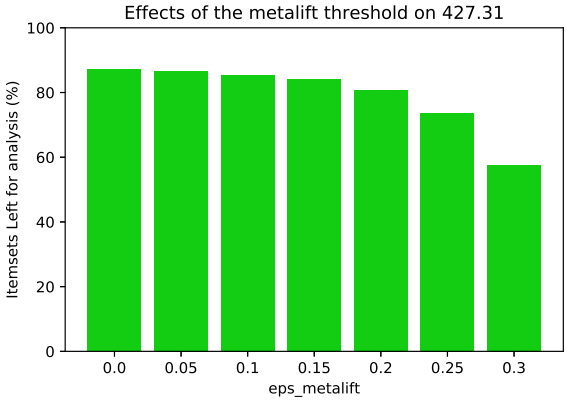
(g) 403.90



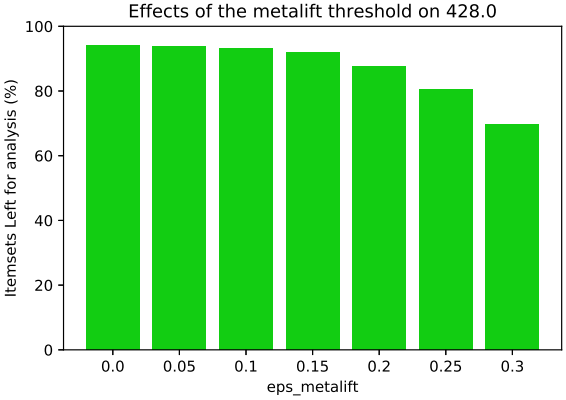
(h) 414.01



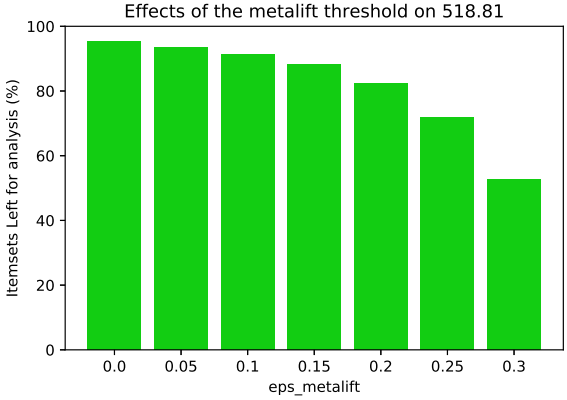
(i) 414.9



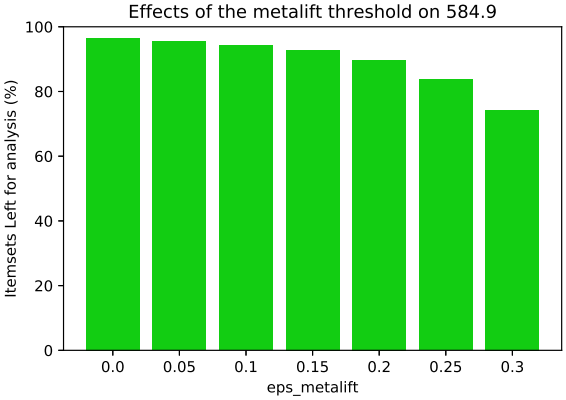
(j) 427.31



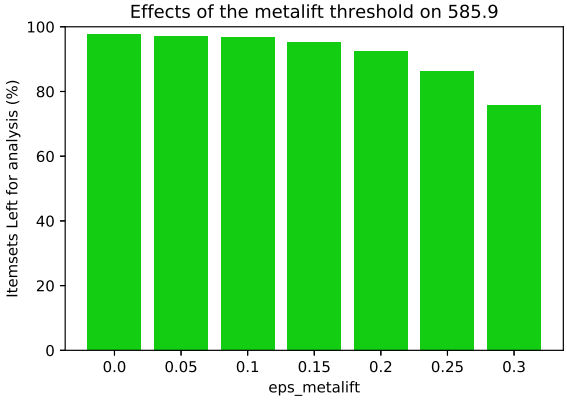
(k) 428.0



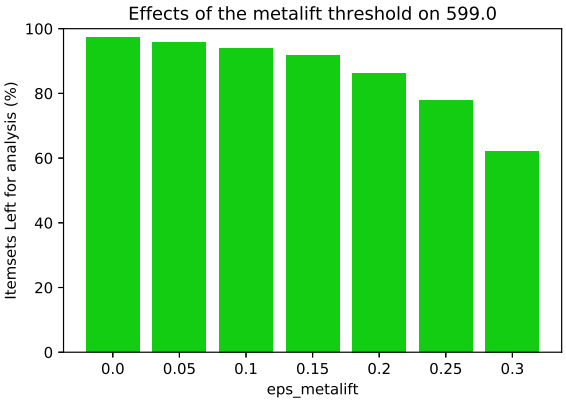
(l) 518.81



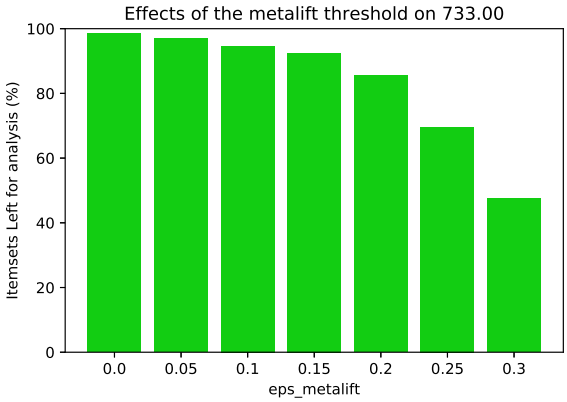
(m) 584.9



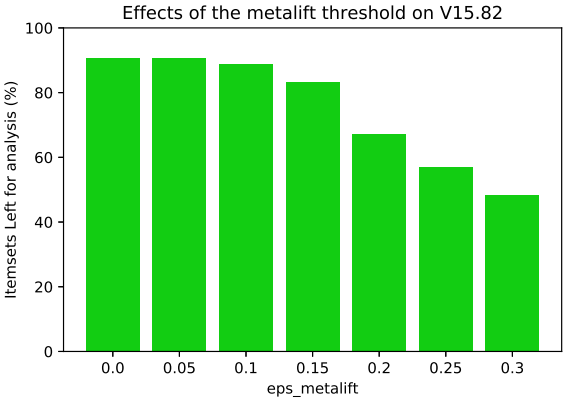
(n) 585.9



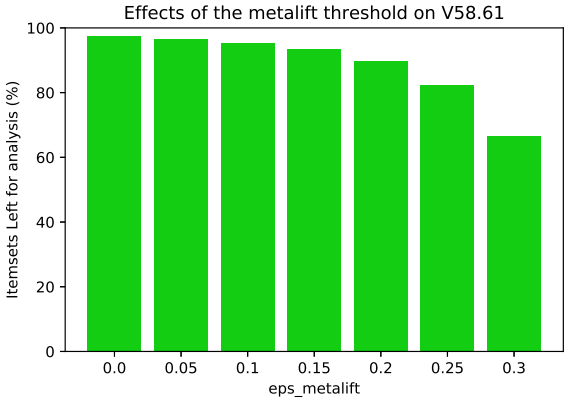
(o) 599.0



(p) 733.00



(q) V15.82



(r) V58.61

Appendix D

Explanatory Variables

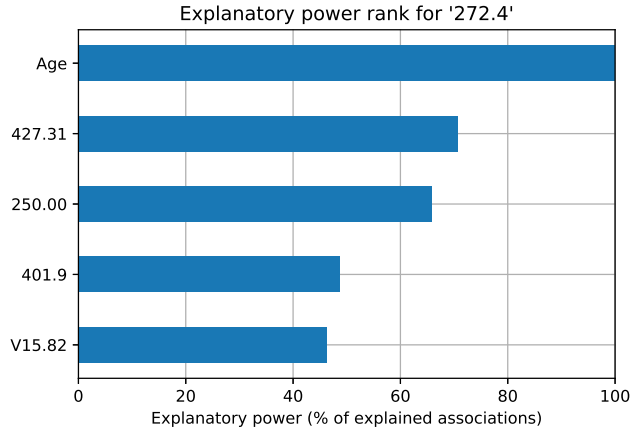
Itemset	**Studied population**	**Metalift**	**Explanatory variable**	**Risk Ratio (*p-value*)**	**Literature**
{041.49, 401.9, Urgent}	250.00	0.69	599.0	61.29 (< 0.0001)	[52, 53]
{294.10, 518.81}	250.00	0.59	427.31	2.54 (< 0.0001)	[54, 50]
{585.9, V58.61, 428.0}	250.00	0.50	427.31	31.94 (< 0.0001)	[55, 49, 56]
{276.7, 584.9}	403.90	0.23	585.9	9.21 (< 0.0001)	[57]
{389.9, 272.4, Urgent}	403.90	0.55	585.9	3.07 (< 0.0001)	[58, 59]
{715.90, 427.31}	403.90	0.53	250.00	2.72 (< 0.0001)	[61, 62]
{496, V58.61, Urgent}	428.0	0.69	427.31	25.53 (< 0.0001)	[63, 49]
{426.3, 250.00}	V15.82	0.63	272.4	5.05 (< 0.0001)	[64, 46]
{585.3, 403.90, 428.0, 272.4, Urgent}	V15.82	0.69	427.31	8.06 (< 0.0001)	[55, 56, 51]
{518.84, 272.4}	V15.82	0.51	250.00	4.65 (< 0.0001)	[65, 46]
{V58.67, V58.66, 250.00, 272.4}	V15.82	0.65	401.9	4.09 (< 0.0001)	[66, 47, 67]
{300.4, 428.0, Urgent}	272.4	0.6	427.31	5.45 (< 0.0001)	[68, 64]
{V45.01, 250.00}	272.4	0.57	427.31	3.99 (< 0.0001)	[69, 62]
{V45.71, V10.3}	311	0.56	Sex=F	228.86 (< 0.0001)	[42]
{365.9, 272.4, 401.9}	250.00	0.58	250.00	3.22 (< 0.0001)	[45, 46, 47]
{V58.61, 518.81, 272.4, Urgent}	403.90	0.53	427.31	34.68 (< 0.0001)	[49, 50, 51]

Some of the most important findings about diagnostics explaining associations are reported above, along with the relative literature confirming our results.

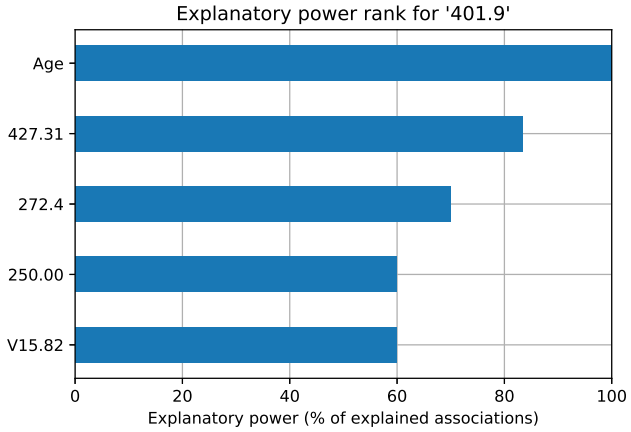
Appendix E

Explanatory Power

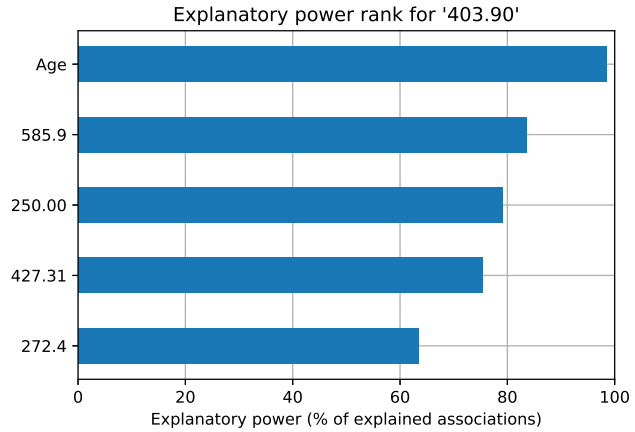
More explanatory power ranks we found interesting are listed below.



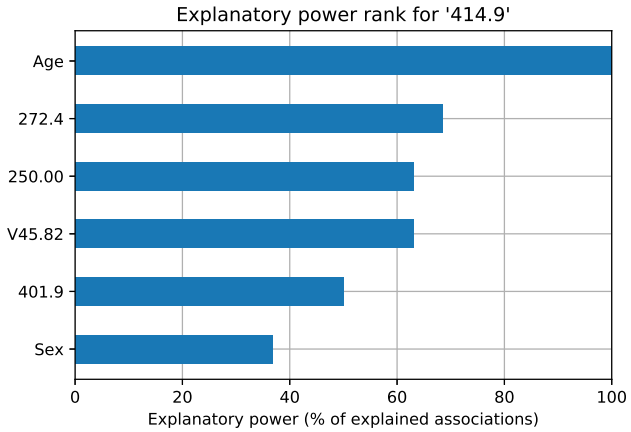
(a) 272.4



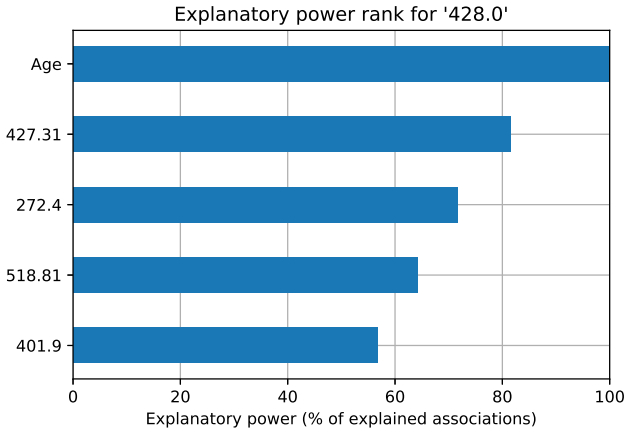
(b) 401.9



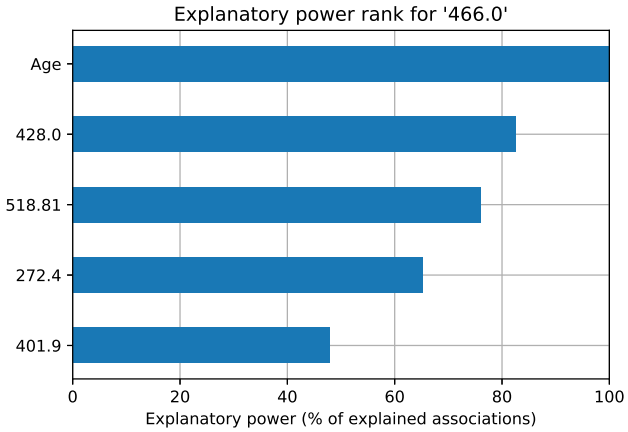
(c) 403.90



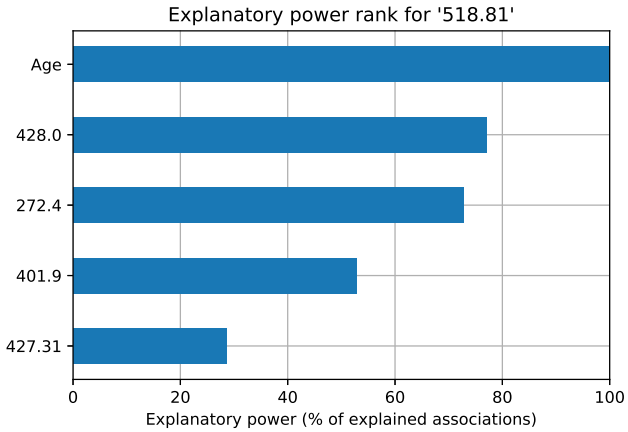
(d) 414.9



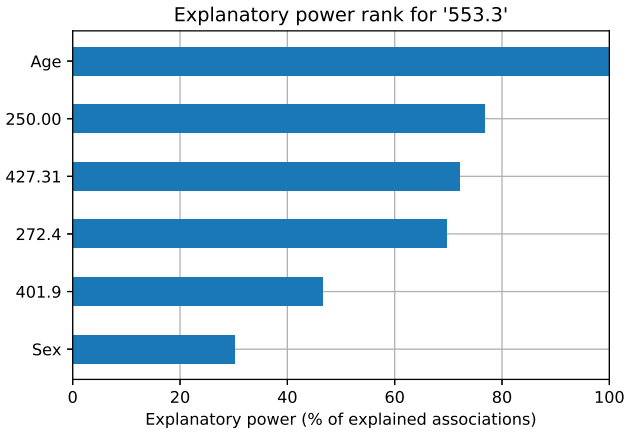
(e) 428.0



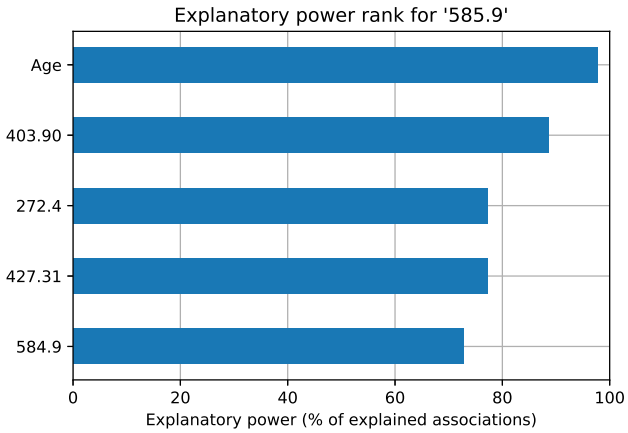
(f) 466.0



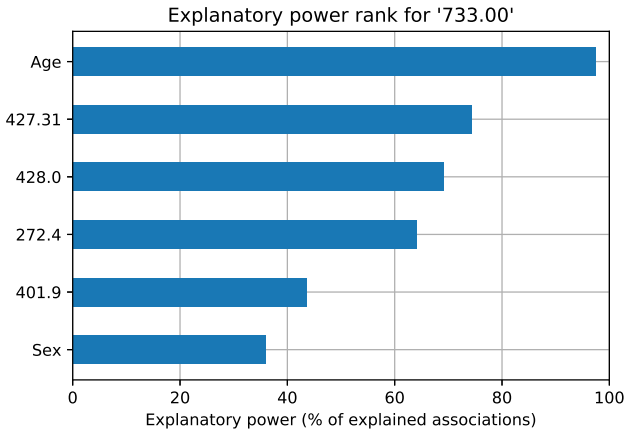
(g) 518.81



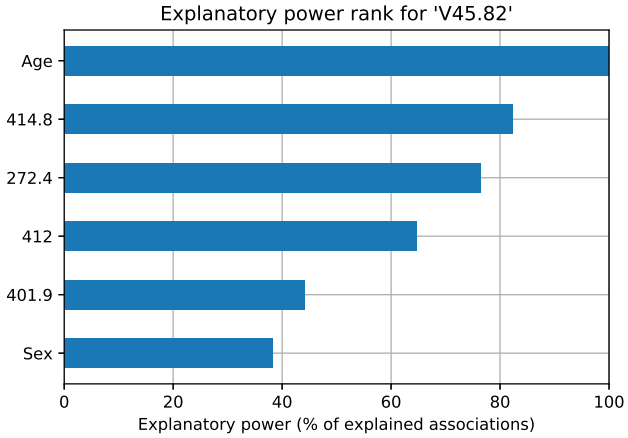
(h) 553.3



(i) 585.9



(j) 733.00


Explanatory power rank for 'V45.82'
Age
414.8
272.4
412
401.9
Sex
0
20
40
60
80
100
Explanatory power (% of explained associations)

(k)V45.82