Characterizing Chronic Disease and Polymedication Prescription Patterns from Electronic Health Records[*]

Martí Zamora[1], Manel Baradad[1], Esther Amado[2], Sílvia Cordomí[2], Esther Limón[3], Juliana Ribera[2], Marta Arias[1], and Ricard Gavaldà[1]

[1]Universitat Politècnica de Catalunya (UPC), Barcelona
[2]Institut Català de la Salut, Barcelona
[3]EAP Mataró-7, Institut Català de la Salut

October 25, 2015

[*] Contact author: R. Gavaldà, gavalda@cs.upc.edu. This paper was presented at the 2015 IEEE Intl. Conf. on Data Science and advanced Analytics, IEEE DSAA 2015, 19-21 October 2015, Paris.

Abstract

Population aging in developed countries brings an increased prevalence of chronic disease and of polymedication - patients with several prescribed types of medication. Attention to chronic, polymedicated patients is a priority for its high cost and the associated risks, and tools for analyzing, understanding, and managing this reality are becoming necessary.

We describe a prototype of a system for discovering, analyzing, and visualizing the co-occurrence of diagnostics, interventions, and medication prescriptions in a large patient database. The final tool is intended to be used both by health managers and planners and for primary care clinicians in direct contact with patients (for example for detecting unusual disease patterns and incorrect or missing medication).

At the core of the analysis module there is a representation of diagnostics and medications as a hypergraph, and the most crucial functionalities rely on hypergraph transversal / variants of association rule discovery methods, with particular emphasis on discovering surprising or alarming combinations. The test database comes from the primary care system in the area of Barcelona for 2013, with over 1.6 million potential patients and almost 20 million diagnostics and prescriptions.

1 Introduction

Healthcare systems are one of the pillars of welfare state. They are also, together with education, one of the largest expenses for public administrations. The fact that most healthcare data is digitalized today opens vast possibilities for improving people's lives and rationalize resource usage: increase our ability to predict and treat disease, select particularly vulnerable individuals for preventive actions, and reduce waste, redundancies, and mistakes.

Many developed countries are experiencing increasing population aging, which implies a large prevalence of chronic disease. Already today, in many western countries, 5% of the population uses up to 50% of the health system resources, uniformly across pharmacy, primary care, and hospitalary care. This makes it even more necessary to optimize healthcare resources from the economic point of view, but also with respect to patient safety and quality of life. A particular concern is the phenomenon of polymedication, i.e. patients that are prescribed several drugs (5, 10, 15, ...), which creates high costs, secondary effects, patient confusion, and low adherence to treatments.

The Catalan Institute of Health (ICS, in short) manages the largest network of ambulatory centers in Catalonia, providing primary health care to about 7 million people, 85% of the Catalan population. One of its strategic lines is innovation in the management of information, including specific programs to exploit electronic health records to improve quality of patient care and sustainability. The vision is that the patient can benefit from a better planning and personalized treatment, and the organization can reduce costs and minimize mistakes, delays, and duplicities.

In this paper we describe the goals and a first prototype of a system designed to allow both health planners and clinicians to analyze, visualize, and put to use healthcare record data from patients, with emphasis on chronic patients and their medication prescription. Health planners could use the tool for understanding the patterns of diagnostics that actually occur in an aging, chronic, polymedicated population, the interaction among different diagnostics and associated medication, temporal trends and geographical and socioeconomic variations. Clinicians at the primary care centers could use the tool for making personalized decisions about their patients, for example by having the tool flag unusual combinations of diagnostics and/or medications (including possible prescription errors), searching for similar cases, or displaying medication suggestions.

The core of the analytical engine of the system is the representation of data as a hypergraph of diagnostics and medications, and hypergraph traversals (variants of itemset and association rule mining), with emphasis on detecting unexpected or alarming patterns. The situation has several peculiarities with respect to standard frequent itemset mining scenarios, among others the existence of two semantically very distinct types of items (diagnostics and medications), of a taxonomy or hierarchy among items, and the amount of unreliable data (most notably, diagnostics that are never deleted even though they are no longer present) that must be repaired somehow.

We developed the tool on the basis of a particular dataset, but hope that it is applicable to many, possibly larger, datasets. It consists of the healthcare annotations (demographics, diagnostics, clinical findings, . . .) and medication prescriptions for the area of Barcelona during 2013, including data from 1.6M patients and almost 20M annotations. We know only of a handful of studies on larger datasets for comparable analysis [9, 12, 14], and either the goals were different or they used the much weaker *graph* (instead of *hypergraph*) representation to represent associations.

2 Context and Project Overview

2.1 Aging, Chronic Disease, and Polymedication

The Catalan Institute of Health (ICS) is responsible for planning and providing healthcare assistance to a large majority of residents of Catalonia, approximately 7 million people. Like in many other European regions, population aging and the increased prevalence of chronic diseases are the largest challenges it faces in the long term. Consequently, one of its priorities is the application of measures for improving chronic patient care, and rationalizing medication for polymedicated patients, favoring the participation of all health professionals and patients themselves in the process.

Chronic diseases are not homogeneously distributed in the population; factors such as geographical area, demographic and socioeconomic structure, country of origin and ethnicity, and the structure of healthcare system itself make it difficult to give uniformly useful indicators. While the volume of available data is already quite large, ICS must be prepared for an immediate future where information collected in real-time and valuation of current scenarios takes precedence over studies of information aggregated in the past months or decades, including subtle trends and cause-effect relations that may escape even the best-trained human experts. It will be necessary to develop powerful and scalable analytical engines that go beyond traditional statistics and can interact with the intuitions and expertise of health professionals.

Polymedication (or polypharmacy) is defined as the use of N or more medications by a patient, where N varies among sources with a minimum of 4. (For reference, patients with 8 and more prescribed medications are not rare in our dataset, and we found one with a staggering 30). It is a result, besides chronicity, of specialization: The cardiologist, the diabetologist, and the hematologist of a single patient will each prescribe their medications, without the time and expertise to study those prescribed by other specialists and investigate possible interactions of redundancies. In addition, the general practitioner does not feel comfortable deleting or modifying the prescription of the specialist. All in all, the consequence is the addition of new drugs to the previous ones, with no critical review of the global medication pattern. A rational, global examination of a patient's list of prescriptions usually leads to alternative lists that are not only cheaper but also safer for the patient. Additionally, studies show that polymedication creates distress and confusion in patients and strongly reduces adherence (i.e., as patients are prescribed more medications, they are more likely to stop taking them). It is thus clear that polymedication puts the patient at risk, sometimes beyond the point of diminishing returns, for a high cost for the system.

The main goal of the project we describe is to build a software system that helps healthcare system managers build maps of diagnostics and prescribed medications in the context of chronic diseases and polypharmacy. This should help better understand a reality that is complex not only clinically but socially, geographically, and economically, and design policies and indicators of their efficiency. Furthermore, some of the knowledge derived from the global panorama could be made available to the clinicians via a dedicated subtool to e.g. offer orientations or raise alerts on specific patients.

In this paper we consider as source of data the clinical history of the patient, which is becoming fully digital across Europe. We focus on structured and semi-structured information, ignoring for the time-being non-structured information such as medical images as it requires a totally different set of analytical techniques. At this stage of the work we are also not concerned with issues such as privacy, security, confederation and interoperability, etc., which will be crucial and blocking if such a tool is to be used in practice.

2.2 The Dataset

The source of data used for the pilot phase of this project has been the electronic clinical history of primary care assistance in the city of Barcelona, and for the year 2013. It can be conceptually described as composed of three kinds of tables:

- A table of recorded *diagnostics* for patients. These include many recorded before 2013 but still considered active, most notably chronic diseases, and refer to diseases, conditions, and clinical findings (such as the results of analytical tests).
- A table of *pharmacy prescriptions,* that is, medications prescribed by primary care doctors to patients during their visits to healthcare centers in 2013.
- Additionally, several small tables with basic demographic information on patients (age, sex, area of residence) and catalogs of diagnostics and pharmaceutical drugs.

We often refer to both *diagnostics* and *pharmacy prescriptions* as *annotations.*

Each patient is associated with a unique code to identify it across tables. The code is an encrypted version of the true patient identifier code used across the health system, created for anonymization before the data was made available to us. Diagnostics and medications are identified by standardized codes known respectively as ICD-10[1] and ATC codes[2]. We do not go into details of the code structure, but it is important to know that they are hierarchical. For example, gastric protector Omeprazol has ATC code A02BC01; code A02BC corresponds to proton pump inhibitors which includes other four active ingredients. In turn, subcodes of A2BC01 might indicate commercial names for the medications containing that active ingredient. ICD-10 codes have formats such as A79.1, where A79 indicates a diagnostic category and .1 a subcategory. They are harder to aggregate than ATC codes because the distinction between category and subcategory is somewhat fuzzy, and what some professionals consider a single disease with variants may go e.g. from

[1] http://en.wikipedia.org/wiki/ICD-10

[2] http://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System

A50 to A64, so the classification must be entered manually; for the moment, we have removed digits after the dot, i.e., have considered categories only.

The data contained 1.6 million potential patients (roughly speaking, residents of Barcelona associated to ICS), 12.3 million diagnostics, and 7.7 million prescriptions. It used originally 8.5 Gb (560 Mb after compression). About 0.5M of the potential patients actually had some associated annotation or prescription - meaning they had visited a primary care center during 2013. We removed approximately 15% of annotations corresponding to patients that did not appear in the main patient table, possibly transient visitors to the area. After processing, we obtained a single table with patient information. For each patient we kept sex, age, residence area, a list of diagnostics and a list of medication prescriptions.

Let us note some limitations of this dataset. First, it corresponds only to primary care at the public network, so it does not include hospital visits and private care centers. This clearly introduces a bias, as areas with different economic levels will rely more or less on private healthcare, and e.g. patients in different age ranges may visit hospitals with different frequency. Secondly, we have data for one year only, which makes it virtually impossible to perform temporal analyses such as evolution of patients or patient types over time; ICS has reliably complete and sorted data since 2009, and we hope to obtain access to it in the near future. Finally, data is entered in the system by people, the health professionals, with procedures that are not completely mechanized and create inconsistencies.

A particular case of inconsistency that we will deal with is what we call *open episodes.* A patient visits his primary care center because of a health condition, which is recorded in the system; the condition disappears after a couple of visits and appropriate medication, so the patient stops going to the center. Yet, the condition remains recorded as one of his/her diagnostics. Even if the patient visits again, perhaps for another reason, the doctor may not think of marking the previous condition as resolved. We have found patients with an arm broken for a year, and women that have been continuously pregnant for over two years. These open episodes are in fact very frequent, and need to be removed so that they do not distort statistics and analyses.

2.3 Project Goals

The long-term goals of the project described here are:

- Contribute to determine the groups of health factors and diagnostics that are most relevant in the population, where relevant depends on prevalence, on impact on patient quality of life, and on impact on the health system.
- Determine the therapeutic areas (treatments, prescriptions) that are most common for a given annotation, and determine their characteristics for better matching with care and attention processes.
- Define and study new indicators of population health that reflects the complexity of current scenarios.
- Define and study new indicators of the impact of health-related plans and programs.
- Facilitate sustainability of health-related processes, optimizing resources and indicating unjustified duplicities.
- Revert the knowledge obtained back to individual patients in order to improve their safety and quality of life.
- Contribute to involving first-line health professionals and patients in this latter goal, by making them more aware of their situation within the general population and of the potential effects of their personal actions.

The specific goal which concerns us is to develop a data analysis tool that basically helps ICS to "map the territory" before making decisions. In particular, we want to build a robust, flexible, and friendly software system that supports the general goals with the following functionalities:

1. Verify the quality and consistency of the data in electronic clinical records.
2. Obtain maps of diagnostics, particularly of chronic diseases with high prevalence, and their relations to patterns of polymedication.
3. Proactively identify anomalous cases of patient medication and diagnostic so that they can be studied by specialists.
4. Find patients that are similar, with respect to the maps built, to a "patient under study".

5. Perform differential analyses according to different criteria, such as demographic or geographic. This should help expert analysis of best practices and comparative analyses of macro- and micro- management.

6. Identify outliers such as districts, health professionals, and centers that have in any sense atypical behaviors or results.

7. Create predictive and explanatory models for variable combinations of interest.

8. Relatedly, project to the future the evolution of variables and populations under study.

9. Identify communities with similar behavior.

Functionalities 3 and 4 are intended for clinicians in direct contact with patients, while the others are more intended for health planners and managers. In this paper we focus on functionalities 1, 2, and 3, as they are the ones which have already been implemented and (partially) validated.

Additionally, the software should satisfy these non-functional requirements:

1. Visually present all of the above interactively so that health professionals with little experience in "big data" tools can use it fruitfully and with little learning time. Users should be able to interactively experiment and select/filter data as they go.

2. Integrate smoothly with the tools that health professionals are already using.

3. Scale up to large volumes of data.

Visualization, integration, and an almost-flat learning curve are completely crucial. Experience shows that many professionals will give a new software at best *one* opportunity. If, on a first trial, they do not immediately find it straightforward to use, consistent with the software they already use, and useful for their packed day-to-day practice, chances are that they will not give it a second chance.

2.4 Related Work

There are of course many studies in knowledge extraction from electronic healthcare records (EHR). See e.g. [15, 19] for general discussions. We focus on those that resemble ours either because the goal is to understand interaction among diseases, or because there is some underlying graph representation.

Some generic studies aim at discovering patterns over the whole patient set without preconceived hypotheses, hoping to find relevant ones that were unnoticed before and suggest other more focused studies. This is the case of [18], which finds new associations between diseases from the EHR of 667,000 patients. In essence, their system find association rules among diseases and medications, like ours, but uses simple rule confidence as a measure.

The study [10] builds networks (graphs) of diseases from the EHR of 327,000 patients, then applies clustering to the networks in order to find those diseases the often occur together in patients; a set of sophisticated heuristics is then applied to highlight the reliable and interesting associations.

The Human Disease Network [9, 12] is the largest study resembling ours that we have found. Roughly speaking, it builds a network (graph) of about a few thousand diseases and genes from a database of over 30M patients. The strength of associations among these items is then studied from the weights of the edges, and the statistics of the network itself are studied with the usual parameters in complex networks studies. The main differences of our work with this one are 1) we consider medication prescriptions instead of "-omics", and 2) we use a hypergraph instead of graph, or in other words, explore relations beyond binary ones. Similarly, the large study [14] used 15 years of registries from the Danish health system, with over 6M patients. Their emphasis is on obtaining temporal trajectories, e.g. evolution of patients and diseases over time (a direction we certainly want to follow in the future); we focus more on the description of associations, the relation to medications, and eventually helping with the definition and evaluation of healthcare policies and personalized patient care.

Many works mine Electronic Health Records with a specific purpose in mind - we mention but a few examples. The study [21] analyzes EHR of 50,000 primary care patients to identify the documentation of the signs and symptoms of heart failure in the years preceding its diagnosis. Mani *et al.* [17] aim at building a method that identifies dementia better than methods existing at the time, using data from EHRs and data mining techniques.

Works such as [11, 22] deal with the detection of adverse drug effects from EHR. Our system is intended to ease studies of the former category, namely, facilitate open-ended exploration of the data by healthcare professionals to gain understanding and intuition which may certainly, among others, suggest in-depth studies of specific questions using other tools.

3 The Prototype: Structure and Functionalities

This section describes the current prototype and typical usage workflows. More details on the most interesting functionalities are given in later sections.

At the core of the prototype there is the representation of a network of associations among diagnostics and prescriptions as a hypergraph. More precisely:

- There is a node for each diagnostic present in the database, and one for each drug; as mentioned before, there is a prior choice of the level of granularity in the hierarchies of diagnostics and drugs.
- An edge between two nodes is labeled by a "weight" computed from the "strength" of the association between the items at the two nodes.
- In fact, this is generalized from pairs of vertices to unordered sets of vertices, denoting the "strength" of the joint association between all the vertices in the set. This turns the structure from a *graph* with binary edges to a *hypergraph* with edges of arbitrary cardinality.
- Alternatively, we use the term *itemset* [1, 2] as synonym of hyperedge. To be precise, only the itemsets corresponding to a certain minimum strength in the database (the *frequent itemsets*) are retained, for efficiency.

The exact definition of "weight" and "strength", and the computation of hyperedges or itemsets will be discussed in Section 4.

The workflow of the prototype is summarized in Figure 1. Initially, all data (diagnostic, prescription, and auxiliary tables) are read from the database or from suitable text files. The frequent itemsets of diagnostics and prescriptions (equivalently, the hypergraph whose vertices are diagnostics and prescriptions) are computed, and the user enters the main menu. At every moment, the options in the main menu apply to a *current population*, for which the itemsets/hyperedges are already computed and which can restricted by successive application of filters. Thus, the current population is initially the whole population read from the DB.

The current options in the main menu of the application are:

- Display statistics on the current population (not shown in Figure 1). These include age / sex descriptors and histograms, diagnostics frequencies and histograms, prescription frequencies and histograms, etc. Many more can be added in the future.

- Filter current population: retain only those that satisfy a boolean query on the variables of interest (currently sex, age, geographical area, diagnostics, and prescriptions). Frequent itemsets (equivalently, the hypergraph) of the new population are computed after the filter.

- Compute, prune, and export rules of interest on the current population, from the itemsets already computed. Details are given in Section 4.

- Navigate the current hypergraph: The user can choose a diagnostic or prescription, a node in the hypergraph, visually inspect the graph around it, and navigate to neighboring nodes. Details are given in Section 5.

Additionally, at any moment the user can save the current population (implicitly, by the sequence of filters that define it) and associated itemsets (hypegraph), or retrieve a previously saved population to continue working on it.

4 Hypergraph Representation of the Data and Rule Computation

As mentioned before, a first approach to understanding the data is to build a network or *graph* where the nodes are annotations (diagnostics or drugs) and edges represent associations between their endpoints. Edges may be annotated with real-valued weights indicating the strength of the association.

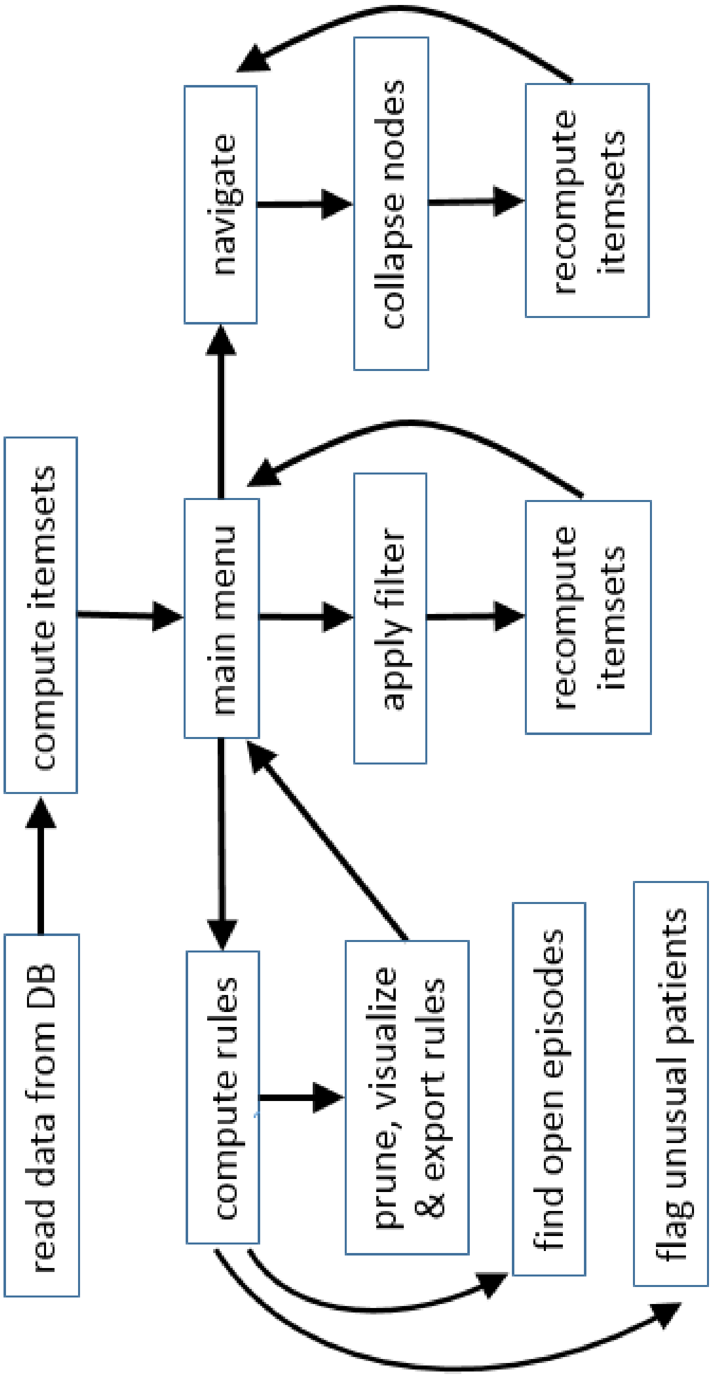


Figure 1: User workflow of the current prototype.

To make the graph sparse and understandable, only edges with weight above a certain threshold (indicating e.g. significant enough association) are kept.

A first definition for the strength of the association between nodes A and B as simply the joint prevalence of A and B relative to the population size. But that can be easily seen to be useless. Some medications such as pain-killers or the already mentioned gastric protector Omeprazol are prescribed very widely and for a large variety of conditions, so they would appear strongly associated to many diagnostics in a very non-specific way. We have considered three measures that are more informative of interesting associations: one based on odds ratios, one based on p-values, and Pointwise Mutual Information (PMI). The latter was finally chosen as it gave most intuitively appealing results. It is defined as:

$$\mathrm{PMI}(A, B) = \log_{10} \frac{\Pr(A, B)}{\Pr(A) \Pr(B)}.$$

Thus, if two items tend to appear together they will have high PMI, items that tend to appear separately will have low (negative) PMI, and items with no particular relation will have PMI close to 0. As an example, if their PMI is 2, we know that the probability of finding them together is 100 times higher than one could expect by chance. This logarithmic scale is useful as without it some values get very large; for example, medication for diabetes is obviously extremely highly associated with diabetes.

Exploring the graph above certainly provides useful information and insights, but also has clear limitations. For example, the joint prevalence of three diagnostics $\{A, B, C\}$ cannot in general be deduced from the joint prevalence of the sets of diagnostics $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$.

For deeper analyses, we move from graphs to *hypergraphs,* where *hyperedges* are arbitrary subsets of items, and this distinguishes our work from many others that remain at the graph level. Note that the definition of PMI above makes perfect sense when A and B are sets of annotations and not single annotations.

To make the process feasible, we will only actually build the hyperedges or subsets that have a minimum support (joint prevalence, in the case of diagnostics) in the population. The problem of hypergraph traversal (or equivalently, finding frequent itemsets) is fortunately well studied and reasonably well solved algorithmically.

The functionalities implemented so far on this hypergraph are:

- Generating itemsets of frequently co-occurring diagnostics.
- Generating association rules mapping diagnostics to medications, and vice-versa.
- Detecting patients with unexplained medication patterns.
- Detecting open episodes, defined in Section 2.2.

For generating frequent itemsets, we used the well-known implementation by Borgelt [5] of the Apriori algorithm [2, 1], which we found to be fast and convenient enough in our context. We did not use standard implementations that compute association rules from frequent itemsets, as we are only interested in particular types of rules. To start with, we want only rules whose antecedent contains only diagnostics and whose consequent is a drug, or else whose antecedent is a single drug and whose consequent is a single diagnostic. Additionally, for the first type of rules clinicians indicated that in most cases the decision to prescribe a drug is made on the basis of a single diagnostic (or at most two), which greatly simplifies the space of association rules to investigate. Note that we still have to compute frequent itemsets of any size for the first functionality desired, finding sets of diagnostics that appear frequently together.

As is often the case, naively applied frequent itemset mining produces a large set of redundant rules. Standard measures of rule interest (support, confidence, leverage, lift, etc.) can to some extend be used to quantify their interest. We filtered more aggressively the set of rules from diagnostics to medication using the following rules:

- Discard a rule $D \to M$ if $\Pr[D, M] \simeq \Pr[D] \cdot \Pr[M]$ (equivalently, if lift is close to 1 or PMI is close to 0). The reason is that when this condition holds there is insufficient evidence that doctors prescribe medication M *because* the patient has diagnostic D.
- Discard a rule $D \to M$ that has confidence σ_1 if there is a rule $D \to D'$ (where D' is also a diagnostic) with confidence σ_2 and a rule $D' \to M$ with confidence σ_3 such that $\sigma_1 \simeq \sigma_2 \cdot \sigma_3$. The intuition is that this way we discard (only) rules that link an infrequent diagnostic with medications that are given for very frequent diagnostics such as diabetes. We have verified that this is the case empirically.

Surely, other heuristics for discarding uninteresting rules can be applied, and some will be suggested by the feedback from experts that examine the outcomes of our system. We are planning to use existing research on principled methods for obtaining non-redundant yet informative subsets of rules [3, 4, 7, 8, 13].

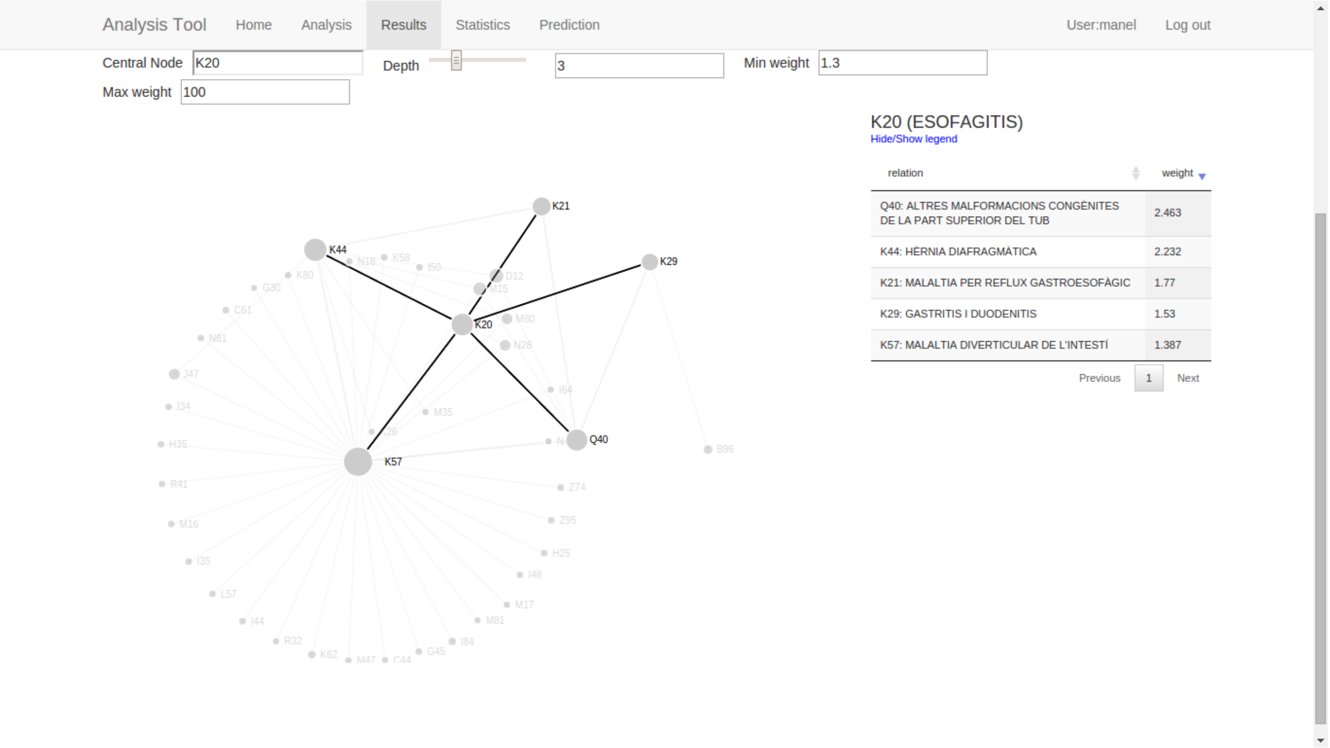
Once we have this set of frequent associations between medications and diagnostics, we can look back at the patient database and look for unexplained diagnostics and prescriptions. If a patient's record contains a diagnostic but none of the medications associated to it, this may indicate either an omission on the clinician's part (a rare, but alarming event) or else that the patient does not have the condition any more, although it remains in his/her record (what we called an open episode). Unfortunately, if applied so naively, this method produces a large number of false positives, cases where the doctor correctly did not prescribe the standard medication. This happens often for chronic conditions such as hypertension, for which doctors may not prescribe the usual medication in moderate cases. Having access to longer historical records from the patient may help in distinguishing between these cases and true open episodes, namely, checking whether some time in the past the patient was actually medicated for the unexplained diagnostic.

Conversely, a drug prescription that is not the consequent of any rule where the antecedent is a set of diagnostics recorded for the patient may indicate either an omission to record a diagnostic or an incorrectly prescribed medication. Both should be flagged as surprising or suspicious when the patient is being prescribed.

5 Visualizing the Hypergraph

From the main menu, the user can also visualize and navigate in the hypergraph of annotations. S/he chooses a node on which to focus (selecting the desired diagnostic or drug), a minimum weight and a "depth", and the graph of neighbors of the chosen node up to that depth or distance and with at least that association weight are shown.

Figure 2 shows an example screenshot with the focus on K20, Esophagitis (diagnostic and drug names show in Catalan), with depth 3 and minimum weight 1.3. Edges from the focus to their immediate neighbors are shown in black, and other edges are lighter so as not to obscure the view. The panel on the right lists immediate neighbor nodes in decreasing PMI order. For



Central Node K20

Depth 3

Min weight 1.3

Max weight 100

K20 (ESOFAGITIS)

relation	weight
Q40: ALTRES MALFORMACIONS CONGÈNITES DE LA PART SUPERIOR DEL TUB	2.463
K44: HÈRNIA DIAFRAGMÀTICA	2.232
K21: MALALTIA PER REFLUX GASTROESOFÀGIC	1.77
K29: GASTRITIS I DUODENITIS	1.53
K57: MALALTIA DIVERTICULAR DE L'INTESTÍ	1.387

Figure 2: Example screenshot: Graph around the Esophagitis diagnostic.

example, one can see that in the current population Esophagitis is most associated to “Other congenital malformations of upper gastrointestinal tract” (PMI 2.463), “Diaphragmatic hernia” (PMI 2.232), etc.

When the focus is on a node, the user may click on any other visible node to move the focus to it.

In order to investigate non-binary associations (i.e., visualize the hypergraph), we have implemented the ability to merge nodes. The user may click on an edge linking two nodes A and B and select to merge them. The nodes are removed and replaced with a new node A-B representing the joint occurrence of A and B. An edge from A-B to another node C has weight $PMI(\{A, B\}, C)$.

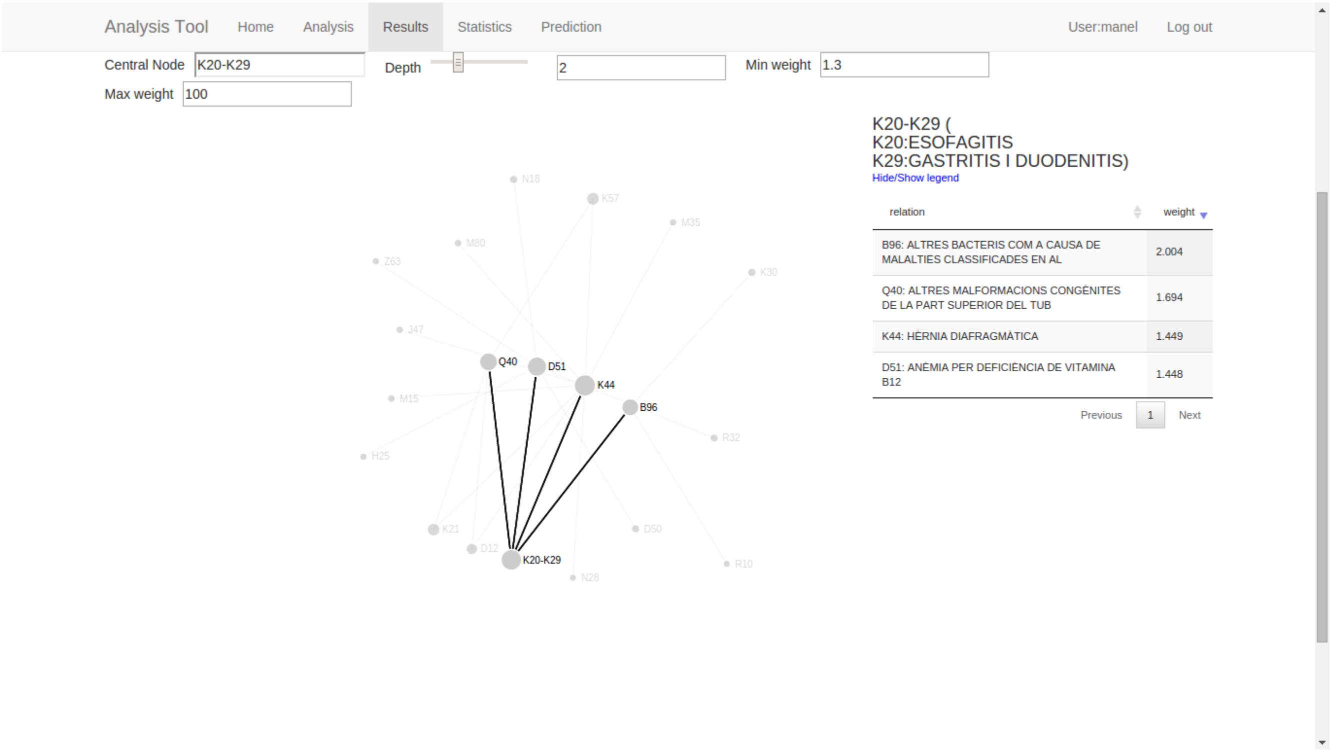
Note that this value cannot be deduced simply from the itemset frequencies for binary associations between A, B, and C that were already computed. A recomputation of frequent itemsets restricted to patients whose annotations contain $\{A, B\}$ is required. This is currently implemented as a sequential scan of the current population to retrieve these patients (which takes 5-10 seconds on our current implementation and the full DB), then a new call to Apriori on this subpopulation (which takes less than 1 second). In a future version we will investigate indices that help efficiently retrieve the patients with $\{A, B\}$ and improve the interactivity of the exploration of large populations.

Figure 3 shows the result of merging diagnostics K20 (Esophagitis) and K29 (Gastritis and Duodenitis) into K20-K29, and Figure 4 shows the result of further merging K20-K29 with Q40 (Other congenital malformations of the upper gastrointestinal tract) into a new node K20-K29-Q40. One can see on the right panel that new associated nodes may appear because of their PMI has increased, such as here D51 - Anemia due to B12 vitamin deficiency.

6 Implementation

The prototype contains a server and a web client. The server contains the actual data and implements the functionalities described; it is implemented in Java. The client lets the user issue commands, including filters for data selection, and present results; it is implemented in Javascript, initially using the *ngraph* library [16] for graph visualization and later using D3.js [6].

As mentioned, we used Borgelt’s implementation [5] of the Apriori algorithm for itemset mining [2, 1], which was fast enough for our case. Even on



Central Node K20-K29

Depth 2

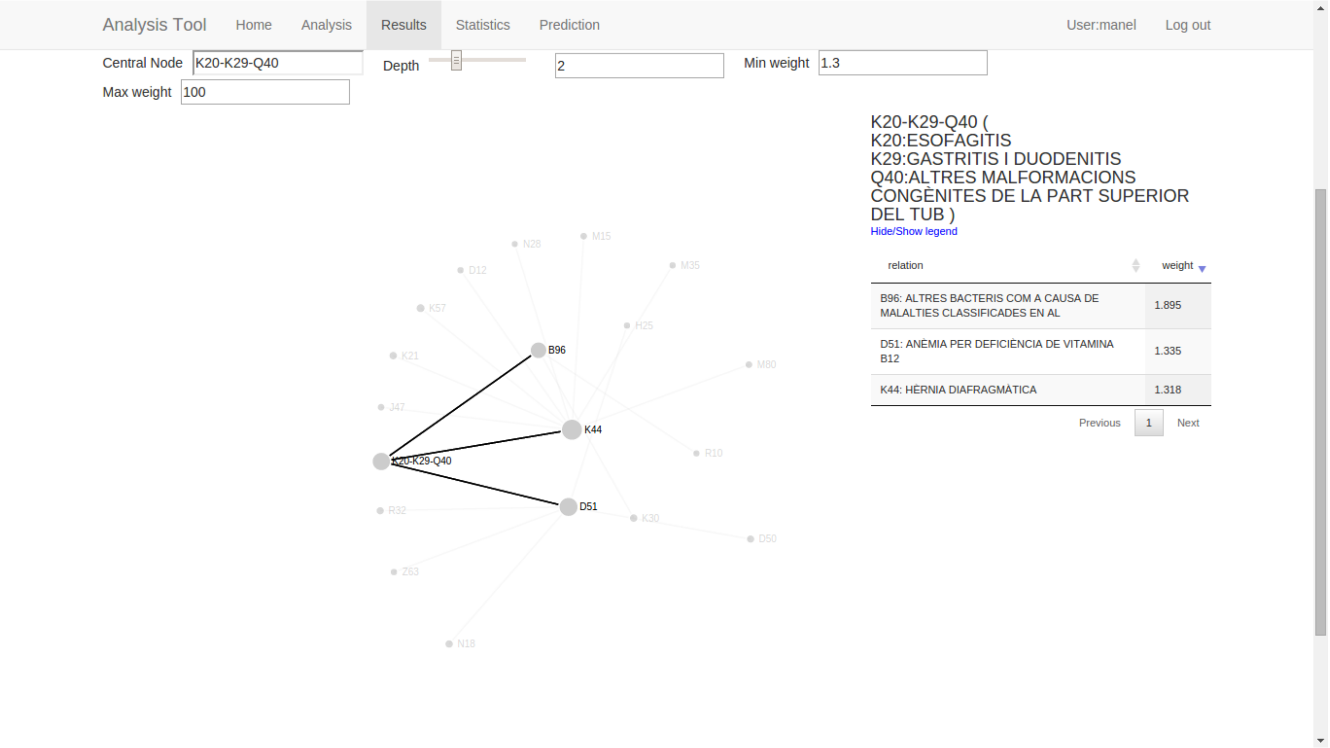
Min weight 1.3

Max weight 100

K20-K29 (
K20:ESOFAGITIS
K29:GASTRITIS I DUODENITIS)

relation	weight
B96: ALTRES BACTERIS COM A CAUSA DE MALALTIES CLASSIFICADES EN AL	2.004
Q40: ALTRES MALFORMACIONS CONGÈNITES DE LA PART SUPERIOR DEL TUB	1.694
K44: HÈRNIA DIAFRAGMÀTICA	1.449
D51: ANÈMIA PER DEFICIÈNCIA DE VITAMINA B12	1.448

Figure 3: Example screenshot: Result after merging K20 (Esophagitis) and K29 (Gastritis and Duodenitis).



Central Node K20-K29-Q40

Max weight 100

Depth 2

Min weight 1.3

K20-K29-Q40 (
K20:ESOFAGITIS
K29:GASTRITIS I DUODENITIS
Q40:ALTRES MALFORMACIONS CONGÈNITES DE LA PART SUPERIOR DEL TUB)

relation	weight
B96: ALTRES BACTERIS COM A CAUSA DE MALALTIES CLASSIFICADES EN AL	1.895
D51: ANÈMIA PER DEFICIÈNCIA DE VITAMINA B12	1.335
K44: HÈRNIA DIAFRAGMÀTICA	1.318

Figure 4: Example screenshot: Result after merging K20-K29 with Q40 (Other congenital malformations of the upper gastrointestinal tract).

the full data and relatively low supports (well below 1%) running time on a standard laptop was in the order of a few minutes. Thus, interactive analysis on subpopulations of interest may be within reach.

The system is designed in three layers: data, logic, and visualization. Two implementations of the data layer are provided now: In one, the data is kept in RAM; in a standard PC, RAM was barely sufficient to hold the current dataset, but would not scale e.g. to a larger population or larger time frame. The other uses the Sparksee graph-oriented database [20]; the graph maintains nodes representing patients, diagnostics, and medications, and edges their relations. Note that the hypergraph we build using Apriori does not contain patient nodes, but the external database does, because the transactions on which Apriori has to be applied are the patients. Observe that patients are the bulk of information of the database; the hypergraph of (only) diagnostics and drugs is likely to be small enough to fit in RAM even for large databases.

7 Preliminary Results

7.1 Execution on the Dataset

We mostly worked with a support of 0.05%, corresponding to a minimum of 800 cases in the general population. This leads to a graph with 918 different diagnostics and 268 different medications. Fixing minimum support to 0.05% and minimum confidence to 10% and after pruning with the heuristics discussed, we get 4051 diagnostic-to-medication rules involving 537 diagnostics and 101 medications, and 2253 medication-to-diagnostic rules involving 201 medications and 115 diagnostics.

In approximately 10% of patients we obtained alerts for prescriptions without diagnostics that explain them from the mined rules. This is actually lower than expected, because the system from which the dataset was extracted does not require the clinician to link a prescription with a diagnostic.

Conversely, in about 16% of patients we found open episodes, health problems with no prescription among those that are strongly associated to them in the mined rules. As expected, these cases mostly correspond to trivial or non-specific health problems (flu, sprained ankle) for which a short-duration prescription is issued; the patient does not return for retest so the diagnostic is not removed from the history.

7.2 Clinical Significance

The preliminary sets of rules obtained are currently being examined by expert clinicians as well as pharmacists. A first observation is that the rules with largest support appear, as was to be expected, highly prevalent diagnostics such as hypertension, diabetes, and obesity. Those with largest confidence relate e.g. diabetes with medication for diabetes, again as expected. Less expected ones link e.g. diabetes with statins (a widely used medication for reducing cholesterol).

A second observation is that, as usual in data mining, the association obtained can be informally grouped in three categories. We give one or two preliminary examples of each:

1. Well known, not surprising, but still reassuring that they were discovered by the program: the association between diabetes and retinopathy, or the fact that that gastric protectors (such as Omeprazol) are prescribed for basically every nontrivial health problem, not because of the health problem itself, but because of the medication prescribed due to it.

2. Unknown and believable ("never thought of it, but it makes perfect sense"): strong association between bedsores (pressure sores) and Alzheimer's disease. The health manager in our team immediately observed that there was no particular prevision for dealing with bedsores in Alzheimer's patients, and that of course there should be one.

3. Unknown and surprising ("never thought of it, and can't think why it came out"): the fact that retinopathy is more strongly associated with hypertension than with diabetes.

The above were obtained on the whole population, so no spectacular discovery can be expected. The hope is that specialists can reveal new and interesting knowledge by exploring subpopulations defined by specific diagnostics or features. Detailed analyses of the clinical findings are out of the scope of this paper and will be published elsewhere.

8 Conclusions and Future Work

Our system, even at this early stage, seems well able to identify the patterns of joint diagnostics and associated drugs. PMI seems like a valid measure for detecting interesting associations between annotations. A deeper validation and analysis of the results is the main work we require from our associated medical experts in the near future. Also, whether the heuristics for detecting suspicious patients (with mismatching medications and diagnostics) and open episodes are sufficient or need further refinement will have to be investigated.

The testing clinicians are satisfied with the tool for detecting inconsistencies. Not surprisingly, they request a friendlier interface, and more validation rules for filtering alerts.

Besides implementing the pending features, several avenues for further research have opened up during this phase of the project. Some include:

- Deal with taxonomies in diagnostics and medications. Taxonomies in frequent itemset mining has received some attention in the past.

- Study patient "trajectories" over time, i.e., how patients and their sets of diagnostics and medications tend to evolve over time [14].

- Study the graph of associations from the point of view of network analysis: investigate powerlaw distributions, small-world phenomena, measures of centrality and influence (betweenness, pagerank), clustering coefficients and transitivity, community structures, etc.

- Incorporate "viewpoints" for clinicians: Display statistics on patients similar to the current one; display courses of actions taken by other clinicians on similar patients; predict possible evolutions of the patients (probability of recovery and of complications), using predictive models learned from past data.

- In a perhaps distant future, when they become available, link these data with genetic data for improved analysis and prediction.

As a final, non-technical, conclusion we remark the need for cross-disciplinary teams in this kind of transversal projects. Our team includes two computer science faculty members working on data mining, two computer science undergraduate students, a clinician (Ph.D.) working at a primary care center with daily contact with patients, a medical doctor specializing in healthcare information systems and project management, a pharmacist (Ph.D.) involved in planning and deployment of prescription policies, and an economist specialized in healthcare economics at ICS.

Acknowledgements

Research at UPC was partially funded by SGR2014-890 (MACDA) project of the Generalitat de Catalunya, MINECO project APCOM (TIN2014-57226-P), and the support of Fundació Barcelona Media to project SESAAME. We are grateful to José Luis del Val at ICS for the extraction and anonymization of data, Sebastià Santaeugènia for insightful comments on the future of the system, and Sparsity Technologies for providing a free research license of Sparksee.

References

[1] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proc. 20th Intl. Conf. on Very Large Data Bases*, pages 487–499, 1994.

[3] Mafruz Zaman Ashrafi, David Taniar, and Kate Smith. Redundant association rules reduction techniques. In Shichao Zhang and Ray Jarvis, editors, *AI 2005: Advances in Artificial Intelligence*, volume 3809 of *Lecture Notes in Computer Science*, pages 254–263. Springer, 2005.

[4] José L. Balcázar. Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science*, 6(2), 2010.

[5] Christian Borgelt. Sofware for frequent pattern mining. `http://www.borgelt.net/software.html`. Last access: 2015-08-22.

[6] Mike Bostock. D3.js: Data-driven documents. `http://d3js.org//`. Last access: 2015-08-22.

[7] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets: A condensed representation of boolean data for the approximation of frequency queries. *Data Min. Knowl. Discov.*, 7(1):5–22, 2003.

[8] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3), 2006.

[9] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[10] David A. Hanauer, Daniel R. Rhodes, and Arul M. Chinnaiyan. Exploring clinical associations using '-omics' based enrichment analyses. *PloS one*, 4(4):e5203+, apr 2009.

[11] Rave Harpaz, Krystl Haerian, Herbert S. Chase, and Carol Friedman. Mining electronic health records for adverse drug effects using regression based methods. In *Proceedings of the 1st ACM International Health Informatics Symposium, IHI'10*, pages 100–107, 2010.

[12] César A. Hidalgo, Nicholas Blumm, Albert-László L. Barabási, and Nicholas A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353+, apr 2009.

[13] Szymon Jaroszewicz and Dan A. Simovici. Pruning redundant association rules using maximum entropy principle. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2002*, pages 135–147, 2002.

[14] Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5, 2014.

[15] P.B. Jensen, L.J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[16] Andrei Kashcha. ngraph, a graph drawing library. `https://github.com/anvaka/ngraph/`. Last access: 2015-08-22.

[17] Subramani Mani, William Rodman Shankle, Michael J. Pazzani, Padhraic Smyth, and Malcolm B. Dick. Differential diagnosis of dementia: A knowledge discovery and data mining (KDD) approach. In *Annual Symposium of the American Medical Informatics Association, AMIA 1997*, 1997.

[18] Irene M. Mullins, Mir S. Siadaty, Jason A. Lyman, Ken Scully, Carleton T. Garrett, W. Greg Miller, Rudy Muller, Barry Robson, Chid Apté, Sholom M. Weiss, Isidore Rigoutsos, Daniel E. Platt, Simona Cohen, and William A. Knaus. Data mining and clinical data repositories: Insights from a 667, 000 patient data set. *Comp. in Bio. and Med.*, 36(12):1351–1377, 2006.

[19] Naren Ramakrishnan, David A. Hanauer, and Benjamin J. Keller. Mining electronic health records. *IEEE Computer*, 43(10):77–81, 2010.

[20] Sparsity Technologies. The Sparksee graph database (formerly known as DEX). `http://www.sparsity-technologies.com/`. Last access: 2015-08-22.

[21] Steven R. Steinhubl, Kenney Ng, Jimeng Sun, Roy J. Byrd, Zahra Daar, Brent A. Williams, Christopher deFilippi, Shahram Ebadollahi, and Walter F. Stewart. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of Cardiac Failure*, 20(7):459–464, 2014.

[22] Gianluca Trifirò, Antoine Pariente, Preciosa M Coloma, Jan Kors, Giovanni Polimeni, Ghada Miremont-Salamé, Maria Antonietta Catania, Francesco Salvo, Anaelle David, Nicholas Moore, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiology and drug safety*, 18(12):1176–1184, 2009.