2017 IEEE 30th International Symposium on Computer-Based Medical Systems

Application of affinity analysis techniques on diagnosis and prescription data

Sanida Theodora
Department of Informatics and Telematics,
Harokopio University of Athens, Greece
Omirou 9, 17778, Tavros, Athens, Greece
sanida.dora@gmail.com

Varlamis Iraklis
Department of Informatics and Telematics,
Harokopio University of Athens, Greece
Omirou 9, 17778, Tavros, Athens, Greece
varlamis@hua.gr

Abstract — **This study performs an Affinity Analysis on diagnosis and prescription data in order to discover co-occurrence relationships among diagnosis and pharmaceutical active ingredients prescribed to different patient groups. The analysis data collected during consecutive visits of 4,473 patients in a 3 years period, focused on patients suffering by hypertension and/or hypercholesterolemia and applied association rule and sequential rule mining techniques. The findings have been validated in the specific dataset using statistical analysis methods.**

Association rule mining shows an association between gastro-oesophageal reflux and the medicines prescribed for hypertension and heart diseases, which agrees with findings in the related literature. Another interesting finding, not yet been reported in related studies is the association between heart diseases, gastroesophageal reflux and insulin-dependent diabetes mellitus for patients that have both hypertension and hypercholesterolemia.

Apart from the medical findings, which must be subject of further research we propose a methodology for the analysis of data collected from a continuous screening process of a group of patients. With the use of data mining techniques we are able to extract and formulate the potential research questions, which are then validated using statistical methods and can also be validated in larger population studies.

Keywords - cholesterol, hypertension, dyslipidaemia, pure hypercholesterolemia, essential (primary) hypertension, affinity analysis, data mining, association rules, sequential patterns.

I. INTRODUCTION

Cardiovascular disease, according to the World Health Organization, is the main cause for the one third of deaths per year, both globally and in Greece. The risk factor for cardiovascular disease and stroke is the combination of hypertension and hypercholesterolemia, resulting in premature mortality and disability [1]. Hypertension is positively correlated to other pathological conditions such as diabetes, cardiovascular disease, thyroid and gastrointestinal diseases [2]. Anxiety and depressive disorders are associated as well [3]. Finally, diabetes, cardiovascular disease, thyroid and kidney failure are positively correlated with hypercholesterolemia too [4].

Affinity Analysis is a data mining technique that can be applied to any process, where agents can be uniquely identified, and information about their activities can be recorded. It is based on the discovery of co-occurrence relationships among activities recorded by the specific individuals and may consider time (or sequence) of activities. It has been quite popular in retail, in the form of market basket analysis in which retailers seek to understand customers' purchase behaviour. When applied in diagnosis and prescription data it may help in detecting frequently co-occurring diagnosis and uncover hidden diagnosis-prescription associations.

The objective of this study is to demonstrate how affinity analysis techniques can be applied to medical data sets in order to generate association rules and sequential patterns. Rules and patterns are related both to the diagnosis and pharmaceutical active ingredients that have been suggested to patients suffering by specific diseases and the early findings can be the basis for a further analysis.

The specific dataset processed in this study contains a large number of patients suffering from hypertension (I10 ICD-10 code) and/or hypercholesterolemia (E78 ICD-10 code). The information recorded during their consecutive visits to the practitioner for a period of three years, includes demographic data, ICD-10 diagnosis codes and prescribed medicines (more specifically the pharmaceutical active ingredients). The analysis of the dataset follows the typical flow of data mining and knowledge discovery, which comprises a) data selection, b) data pre-processing, c) transformation, d) data mining and e) interpretation.

The following section summarizes related work on the application of data mining techniques on hypertension and cholesterol patients' data. Section III details on the proposed methodology for the analysis of diagnosis and prescription data and the tools employed in this study and Section IV illustrates the results from the analysis performed. Finally, Section V summarizes the findings and provides insights for further research.

II. RELATED WORK

Hypertension and increased cholesterol levels are two major reasons behind Coronary Artery Disease and stroke incidents. These two diseases have been the subject of several studies that apply data mining techniques for the analysis of patient data.

The classification of participants to healthy or patient is a major task in related works. In [5] authors studied the medical profile of 694 persons (452 patient and 242 control subjects) using nine different classification models (3 decision trees, 4 statistical models and 2 neural networks)

2372-9198/17 $31.00 © 2017 IEEE
DOI 10.1109/CBMS.2017.114

for predicting hypertension. They have employed the following features: age, gender, family history of hypertension, smoking habits, triglyceride, lipoprotein-A, uric acid and cholesterol levels in blood sample and body mass index (BMI). The Neural network methods (Radial Basis Function networks and Multi-Layer Perceptrons) outperformed all other models in the prediction of hypertension, whereas tree-based models had the worst performance.

An analysis on hypertension data collected for a group of 1,761 patients [6] revealed that the major risk factors of hypertension are in order of importance: family history in hypertension, the waist-to-hip circumference ratio, the level of cholesterol and triglyceride in blood, the BMI, the gender and the consumption of saturated fats. A classification model based on decision trees had the best reported accuracy at 91.1%.

In [7] authors performed a multi-disease analysis on a dataset that included common risk factor data for 2,048 people (154 hypertensive patients, 579 hyperlipidemic and 187 subjects that suffer from both). They trained several classification models (regression, discriminant analysis, etc.) for predicting hypertension and hyperlipidemia and detecting the most informative factors (risk factors). Systolic blood pressure, triglycerides, uric acid, glutamic pyruvic transaminase and gender were at the top positions.

The extraction of association rules from the dataset was the main task in [8], a study on 492 persons (311 patients and 81 healthy), using total cholesterol, LDL, triglycerides, HDL and VLDL as input parameters and the person status (hyperlipidemia patient/healthy) as output. The extracted rules were totally validated by the domain experts and when employed on a rule based classifier gave an impressive 100% accuracy. Another example of the use of sequential patterns in medicine was presented in [9]. The application of sequential pattern mining algorithms on a dataset of 161,497 patients who had been prescribed at least one diabetes medication, allowed the prediction of the next prescribed medications (at the detail level of drug class) within three attempts in the 90% of cases.

The current study was performed on a larger dataset than the majority of the aforementioned approaches and employs for the first time, patient history information both in diagnosis and prescription in an attempt to uncover hidden relations between the two that worth further study.

III. Methodology

The proposed methodology for extracting useful and previously unknown knowledge from the specific dataset follows the steps of the Data Mining process and demonstrates how Association Rule Mining can be a valuable tool for the analysis of symptom and prescription data for groups of patients monitored for a period of time.

A. Data selection

Data are retrieved from a general practitioner's patient record. For preserving the anonymity of patients, only age and gender information have been selected from patients' demographic profile. From the information registered for the different visits of each patient, we kept the prescription and diagnosis information and the date of the visit. The selected dataset spans the period from January 2012 to July 2015, and contains anonymous data for 4,473 patients located in the wider area of Athens, Greece.

A first analysis of the dataset revealed that the majority of patients in the dataset was diagnosed, at least once, with hypertension or hypercholesterolemia. Since we are interested to see the diseases that are frequently associated with these two, and the pharmaceutical substances that are usually prescribed in those patients, we focused our analysis to the following patient sub-groups:

- Those who have been diagnosed with hypertension at least in one visit. This dataset (HT) includes 1,762 distinct patients (1,028 women and 734 men) and 10,136 distinct visit records.
- Those who have been with hypercholesterolemia at least once. This dataset (HCL) includes 1,295 unique patients (805 women and 490 men) and 7,455 distinct visit records.
- Those who have been diagnosed with hypertension and hypercholesterolemia at least once. This dataset (HCT) contained 729 patients (448 women and 281 men) and 5,370 distinct visit records.

The patients' age and gender, a unique patient ID, the visit or prescription date, the diagnosis ICD-10 codes, and the prescribed active pharmaceutical substance have been recorded in all datasets. Any other information referring to suggested medical exams (e.g. x-rays, CT, MRI, chemotherapy, radiation, blood and urine exams) was removed from the dataset.

B. Data pre-processing

The most important tasks of this step were the identification of patient visits and the extraction of any potentially useful information from textual descriptions.

It is important to note, that a patient may report multiple diseases in the same visit and consequently been prescribed a set of medicines. Because of a restriction to the number of substances that can be given in the same prescription (maximum three) multiple prescriptions may be recorded for the same patient visit. In the pre-processing step, this information was merged and the final dataset contained data for 17,758 distinct visits of 4,154 distinct patients.

In each visit one or more diseases have been recorder using ICD-10 codes and one or more substances have been prescribed. When no ICD-10 codes or medical substance names were given, a string matching technique and human intervention were used to extract as much information as possible from the textual description.

C. Data transformation

The data transformation step prepares the dataset for being used as input to the data mining algorithm. Transformations (e.g. normalization, discretization of continuous values, feature composition etc) may improve the performance of the data mining algorithm or diversify the type of knowledge extracted from the same dataset. For association rule mining it is important that the input dataset has the form of a transaction database where each transaction contains one or more items or activities. When the order of items in the transaction is not important, the aim of the association rule algorithm is to locate frequently co-occurring items (i.e. items that appear many times together in transaction sets), called frequent *item-sets*. When the order of items is important, we refer to *item-sequences*, which are the basis for sequential association rules (i.e. items that frequently occur in transactions, always with the same order). Typically, the intensity (or weight or importance) of items in a set is not important for the association rule, so duplicate items inside the same set are ignored.

A transformation that may lead to different type of knowledge extracted from the dataset is the aggregation of multiple transactions to larger transactions. For example, before any transformation, from the data it is possible to locate diseases that co-occur frequently among visits, or drugs that are usually prescribed together. However, when the consecutive visits of the same patient are aggregated then a different transaction database is created, where each patient is a transaction, the list of diagnosis contains every distinct diagnosis for the patient and the prescription column contains all the distinct drugs that have been prescribed to the patient. The knowledge extracted from the aggregated transaction database will have the form of diseases that frequently co-occur in the patients of the practitioner or sets of medicines that many of his patients have taken at least once in the examination period.

D. Data mining algorithms

The extraction of *association rules* is one of the most important data mining techniques, which aims in discovering interesting relations among items based on their frequent co-occurrence in user transactions [10].

An association rule is in the form:

$$X \Rightarrow Y$$

where $X, Y \subseteq I$ and I is the set of all items. The first step for the extraction of association rules is the extraction of frequent item-sets (i.e. sets that comprise all items in $X \cup Y$).

The rules are evaluated for their importance, significance and interestingness using a long list of evaluation metrics. The most popular among these metrics [11] are:

- Support: the ratio of transactions that contain all items of the rule ($X \cup Y$) to the number of transactions in the database.
- Confidence: the proportion of the transactions that contain the items in the left hand side (LHS) of the rule (i.e. X), which also contain the items in the right hand side (RHS) of the rule (i.e. Y).
- Lift: the ratio of the observed support to that expected if items in both sides were independent.

FP-Growth is a fast association rule mining algorithm [12] that has been employed in this study for extracting frequent item-set. As far as it concerns frequent item-sequence extraction, the Generalized Sequential Patterns (GSP) algorithm has been used [13]. The association and sequential rules were extracted from the datasets using a minimum support and minimum confidence threshold in the two algorithms, which limits the number of generated rules and increases their importance.

All the aforementioned steps and the associated tasks have been implemented on RapidMiner Studio, a popular data mining software that allows the definition of a processing pipeline, which leads from the raw data to the extracted knowledge using a rich library of parametric modules.

E. Interpretation of the results

Before interpreting the results of the data mining algorithms it is important to have a better understanding of the dataset. It is also important to present the general characteristics of the sample of study, so that the findings can be comparable to a future study on a similar sample.

The analysis of the dataset focuses in the distribution of cases per disease and the list of the most popular diseases. It also comprises the distribution of cases per age group[1] and gender for the most popular diagnoses and the most prescribed medicaments.

It is also important to evaluate the usefulness and novelty of the extracted knowledge. For this reason, support and confidence thresholds have been employed. Since the set of rules was still large and contained many redundant rules (i.e. rules conveying the same information), only the rules created from the maximal closed frequent item-sets [14] were kept. The extracted rules were ranked in decreasing lift order and manually evaluated by a domain expert.

The rules generated from the two data mining algorithms have one of the following forms:

0.414: <Diagnosis_E78, Diagnosis_I10>
[Substance_AMLODIPINE] --> [Substance_VALSARTAN] (confidence: 0.987)

where the first one shows a frequent item-set and the second one a rule based on another frequent item-set. The number in the first row denotes the support of the item-set and the number in the second row the confidence of the rule. The order of items in the second case denotes the antecedent or left-hand-side (LHS) and consequent or right-hand-side (RHS) of a rule.

[1] The World Health Organization's classification of age groups has been employed.

IV. Results

Before applying the data mining algorithms to each subset of patients, it is important to gain a better insight on the distribution of patients by age and gender in each group and project it to large population size. Figure 1 contains the population percentage per gender and group age for the whole set of patients (top chart) and the three subgroups (patients with hypertension/I10, hypercholesterolemia/E78 and both/I10&E78). The rates have been projected to 100,000 persons, using the Greek Statistics Authority (GSA) data for the population of the city of Athens, which had 3.8 million inhabitants according to the 2011 census.

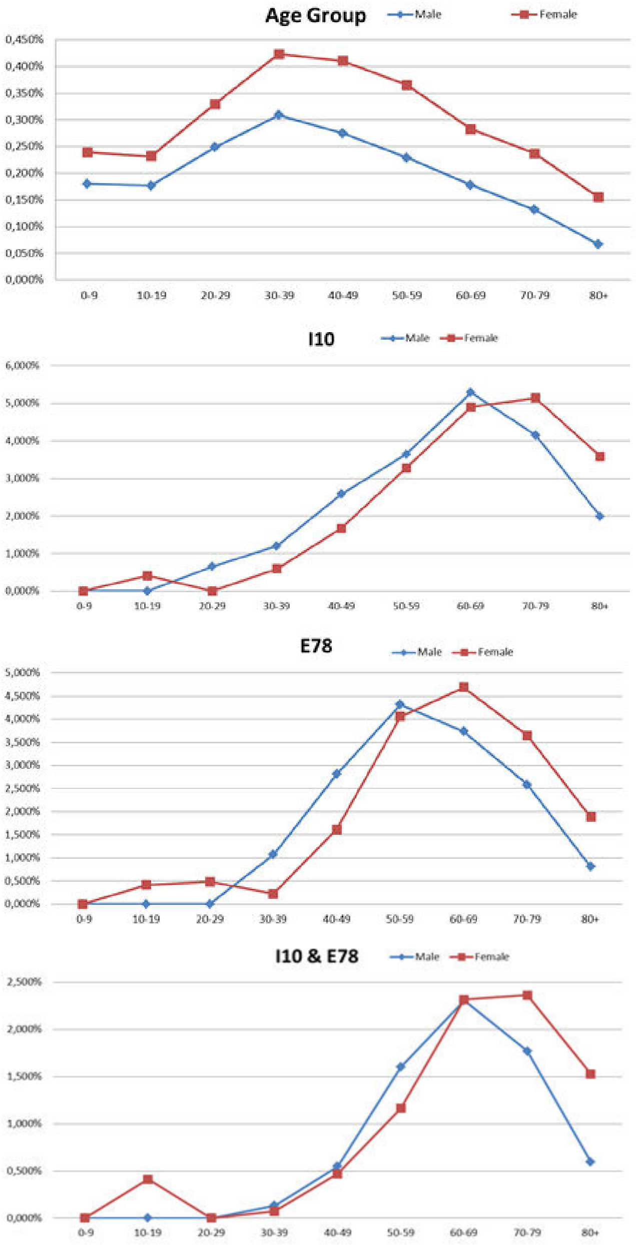

Figure 1. The distribution of patients per age and gender for the whole group (top) and the sub groups. Allocation is done in 100,000.

Results in Figure 1 show that in the study group, women suffer of hypertension and hypercholesterolemia at a later (age) stage than men. The majority of women patients falls in the 60-69 group whereas men's majority falls in the 50-59 one. In the group that has both diseases the age group of 70-79 is the majority among women and 60-69 among men.

The next step of the analysis focuses on the distribution of diagnosis across patients. Figure 2 presents the projected (to 100,000 population) number of patients that have been diagnosed with each of the top most frequent diseases in the dataset. Hypertension and hypercholesterolemia are the most popular, with Gastro-oesophageal reflux (K21) and upper respiratory infections (J06) to follow. Insulin-dependent diabetes mellitus (E10) and Ischaemic cardiomyopathy (I25.5) rank lower.

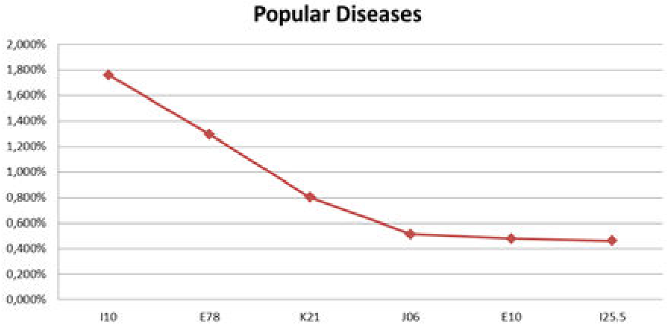

Figure 2. The six most popular diseases in the dataset

Since the dataset combines disease and pharmaceutical substance information, it is important to know the substances that are associated (prescribed) with each disease. Table I provides the main substances subscribed for each disease.

TABLE I. The pharmaceutical substances prescribed for the top diseases in the set

Diagnosis	Substance	Diagnosis	Substance
E78	Atorvastatin	I10	Hydrochlorothiazide
	Simvastatin		Valsartan
	Rosuvastatin		Irbesartan
	Pitavastatin		Olmesartan
	Ezetimibe		Amlodipine
E10	Metformin	I25.5	Clopidogrel
K21	Omeprazole		

The dataset analysis was followed by the execution of FP-Growth and GSP algorithms and the evaluation of results. In FP-growth a minimum support of 10% has been used and a minimum confidence of 70%. Only the maximal closed frequent item-sets have been used for rule generation. In GSP, lift has been used for ranking patterns with support larger than 10%.The three subsets have been processed at visit level (each visit is a transaction) and at patient level (each patient is considered as a transaction).

In the following we refer to the datasets using the following short names:

- HCL stands for the Hypercholesterolemia subset. HCLv when a visit is considered a transaction, HCLp when all visits are aggregated per patient.
- HT stands for Hypertension subset and HTv and HTp are used respectively.
- HCT refers to the group of patients that have both diagnoses. HCTv and HCTp as used as above.

Since the number of association rules that pass the confidence and support thresholds is still large, we provide a summary of the associations and patterns extracted by the two algorithms for the different subsets. A tick (√) in the respective column denotes that in the specific group of patients there is an association between.

TABLE II. ASSOCIATIONS AND PATTERNS PER PATIENT THROUGHOUT ALL VISITS

	Association Rules			Sequential Patterns		
Diagnosis	**Dataset**					
	HCLp	*HTp*	*HCTp*	*HCLp*	*HTp*	*HCTp*
Hypertension	√			√		
Hypercholesterolemia		√			√	
Heart disease	√		√	√	√	√
Gastroesophageal reflux disease	√		√	√	√	√
Insulin-dependent diabetes mellitus	√	√	√	√	√	√
Non-insulin dependent diabetes mellitus						√
Anxiety disorders					√	√
Depression						√
Osteoporosis						√

TABLE III. ASSOCIATIONS AND PATTERNS PER VISIT

	Association Rules			Sequential Patterns		
Diagnosis	**Dataset**					
	HCLv	*HTv*	*HCTv*	*HCLv*	*HTv*	*HCTv*
Hypertension	√			√		
Hypercholesterolemia		√			√	
Heart disease			√	√	√	√
Gastroesophageal reflux disease			√	√	√	√
Insulin-dependent diabetes mellitus					√	√
Non-insulin dependent diabetes mellitus						
Anxiety disorders					√	√
Depression						√
Osteoporosis						

Results in Tables II and III show that for patients suffering from hypercholesterolemia (HCL), there is a positive association with hypertension, coronary artery disease, gastroesophageal reflux disease and insulin dependent diabetes mellitus. In patients suffering from hypertension (HT) there is a positive correlation with hypercholesterolemia, gastroesophageal reflux disease, coronary heart disease, insulin dependent diabetes and anxiety disorders. In patients suffering from both (HCT), there is a positive correlation with coronary artery disease, gastroesophageal reflux disease, insulin dependent diabetes mellitus, non-insulin dependent diabetes, anxiety disorders, depressive disorders, and osteoporosis.

A. *Analysis*

Of the 4,473 patients in the database 1,762 have hypertension, 1,295 have hypercholesterolemia and 729 patients have both diseases. We performed an analysis to observe the different distribution of diseases between hypertension and control group (Tables IV - VI).

TABLE IV. HYPERCHOLESTEROLEMIA ANALYSIS AND CONTROL GROUP

	Control Group – 3,178 patients	**Dataset HCL – 1,295 patients**
Hypertension	32,50%	56,29%
Heart disease	7,93%	16,22%
Gastroesophageal reflux disease	15,45%	23,94%
Insulin-dependent diabetes mellitus	9,00%	14,98%

TABLE V. HYPERTENSION ANALYSIS AND CONTROL GROUP

	Control Group – 2,711 patients	**Dataset HT – 1,762 patients**
Hypercholesterolemia	20,88%	41,37%
Heart disease	7,23%	15,10%
Gastroesophageal reflux disease	13,80%	24,23%
Insulin-dependent diabetes mellitus	7,41%	15,83%
Anxiety disorders	3,43%	10,56%

TABLE VI. HYPERTENSION AND HYPERCHOLESTEROLEMIA ANALYSIS AND CONTROL GROUP

	Control Group – 3,744 patients	**Dataset HCT – 729 patients**
Heart disease	8,39%	20,30%
Gastroesophageal reflux disease	15,38%	30,86%
Insulin-dependent diabetes mellitus	9,00%	19,62%
Non-insulin dependent diabetes mellitus	3,18%	10,97%
Anxiety disorders	5,13%	11,93%
Depression	6,01%	11,52%
Osteoporosis	5,21%	10,97%

Table VII displays results relating to the popular active substances per patient throughout all visits and Table VIII show the results concerning popular active substances by date patient visit.

TABLE VII. RESULTS POPULAR ACTIVE SUBSTANCES PER PATIENT THROUGHOUT ALL VISITS

Dataset	Association Rules	Sequential Patterns
HCLp	Amlodipine, Ezetimibe, Hydrochlorothiazide, Irbesartan, Metformin, Olmesartan, Rosuvastatin, Simvastatin, Valsartan.	Amlodipine, Atorvastatin, Hydrochlorothiazide, Irbesartan, Metformin, Rosuvastatin, Simvastatin, Valsartan.
HTp	Amlodipine, Atorvastatin, Hydrochlorothiazide, Irbesartan, Metformin, Olmesartan, Simvastatin, Valsartan.	Amlodipine, Atorvastatin, Hydrochlorothiazide, Irbesartan, Metformin, Olmesartan, Rosuvastatin, Simvastatin, Valsartan.
HCTp	Amlodipine, Atorvastatin, Clopidogrel, Ezetimibe, Hydrochlorothiazide, Irbesartan, Metformin, Olmesartan, Omeprazole, Rosuvastatin, Simvastatin, Valsartan.	Amlodipine, Atorvastatin, Clopidogrel, Ezetimibe, Hydrochlorothiazide, Irbesartan, Metformin, Olmesartan, Omeprazole, Rosuvastatin, Simvastatin, Valsartan.

TABLE VIII. RESULTS POPULAR ACTIVE SUBSTANCES FROM PATIENTS BY DATE VISIT

Dataset	Association Rules	Sequential Patterns
HCLv	Atorvastatin, Hydrochlorothiazide, Simvastatin, Valsartan.	Atorvastatin, Hydrochlorothiazide, Simvastatin, Valsartan.
HTv	Amlodipine, Atorvastatin, Hydrochlorothiazide, Irbesartan, Olmesartan, Simvastatin, Valsartan.	Atorvastatin, Hydrochlorothiazide, Metformin, Omeprazole, Simvastatin.
HCTv	Amlodipine, Atorvastatin, Clopidogrel, Hydrochlorothiazide, Irbesartan, Metformin, Simvastatin, Valsartan.	Atorvastatin, Clopidogrel, Hydrochlorothiazide, Omeprazole, Simvastatin, Valsartan.

On Tables VII and VIII we can see that the data set of patients suffering from hypercholesterolemia, show the active substances amlodipine, atorvastatin, hydrochlorothiazide, irbesartan, metformin, rosuvastatin, simvastatin, and valsartan. In data set from patients suffering from hypertension appear the active substances amlodipine, atorvastatin, hydrochlorothiazide, irbesartan, metformin, olmesartan, simvastatin and valsartan. In data set with patients suffering from both hypertension and hypercholesterolemia simultaneously, the active substances amlodipine, atorvastatin, clopidogrel, ezetimibe, hydrochlorothiazide, irbesartan, metformin, olmesartan, omeprazole, rosuvastatin, simvastatin, and valsartan are present.

V. CONCLUSIONS

In this study, we applied two popular association rule mining algorithms, FP-Growth and GSP, to the diagnosis and prescription data of the patients of a general practitioner. Results uncovered a correlation of hypertension and hypercholesterolemia with coronary artery disease, gastroesophageal reflux disease and insulin dependent diabetes mellitus. Additionally, in patients with hypertension we have observed positive correlation with anxiety disorders. Finally, in patients diagnosed both with hypertension and hypercholesterolemia, a correlation was observed with non-insulin dependent diabetes, anxiety disorders, depressive disorders, and osteoporosis. Finally, the comparison between considering each visit as a separate transaction and the whole patient history as a transaction revealed better associations (with higher support and confidence) in the second case.

Since the dataset comprises the patients of a single general practitioner it is not safe to generalize our findings, and further research with larger data sets is necessary to verify the comorbidity of hypertension, hypercholesterolemia and both conditions simultaneously as well.

REFERENCES

[1] World Health Organization, & UNAIDS. (2007). Prevention of cardiovascular disease. World Health Organization.

[2] Patten, S. B., Beck, C. A., Kassam, A., & Williams, J. V. (2005). Long-term medical conditions and major depression: strength of association for specific conditions in the general population. Canadian Journal of Psychiatry, 50(4), 195.

[3] Carroll, D., Phillips, A. C., Gale, C. R., & Batty, G. D. (2010). Generalized anxiety and major depressive disorders, their comorbidity and hypertension in middle-aged men. Psychosomatic medicine, 72(1), 16-19.

[4] Bhatnagar, D., Soran, H., & Durrington, P. N. (2008). Hypercholesterolemia and its management. BMJ, 337.

[5] Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications, 29(3), 583-588.

[6] Akdag, B., Fenkci, S., Degirmencioglu, S., Rota, S., Sermez, Y., & Camdeviren, H. (2006). Determination of risk factors for hypertension through the classification tree method. Advances in therapy, 23(6), 885-892.

[7] Chang, C. D., Wang, C. C., & Jiang, B. C. (2011). Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert systems with applications*, 38(5), 5507-5513.

[8] Dogan, S., & Turkoglu, I. (2008). Diagnosing hyperlipidemia using association rules. Mathematical and Computational Applications, 13(3), 193-202.

[9] Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. Journal of biomedical informatics, 53, 73-80.

[10] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

[11] Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, *38*(3), 9.

[12] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

[13] Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on* (pp. 3-14). IEEE.

[14] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., & Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In Computational Logic—CL 2000 (pp. 972-986). Springer Berlin Heidelberg.