MeTA: Characterization of medical treatments at different abstraction levels

Personalized tag recommendation based on generalized rules

Elena Baralis, Luca Cagliero*, Tania Cerquitelli, Silvia Chiusano, Paolo Garza

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Dario Antonelli, Giulia Bruno

Dipartimento di Ingegneria Gestionale e della Produzione, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Naeem A. Mahoto

Department of Software Engineering, Mehran University of Engineering and Technology, Indus Hwy, Jamshoro 76062, Pakistan

Abstract

Physicians and healthcare organizations always collect large amounts of data during patient care. These large and high-dimensional datasets are usually characterized by an inherent sparseness. Hence, the analysis of these datasets to figure out interesting and hidden knowledge is a challenging task.

This paper proposes a new data mining framework based on generalized association rules to discover multiple-level correlations among patient data.

*Corresponding author. Tel.: +39 0110907084 Fax: +39 0110907099.

Email addresses: elena.baralis@polito.it (Elena Baralis), luca.cagliero@polito.it (Luca Cagliero), tania.cerquitelli@polito.it (Tania Cerquitelli), silvia.chiusano@polito.it (Silvia Chiusano), paolo.garza@polito.it (Paolo Garza), dario.antonelli@polito.it (Dario Antonelli), alessandro.fiori@ircc.it (Giulia Bruno), naeem.mahoto@faculty.muet.edu.pk (Naeem A. Mahoto)

Preprint submitted to

Specifically, correlations among prescribed examinations, drugs, and patient profiles are discovered and analyzed at different abstraction levels. The rule extraction process is driven by a taxonomy to generalize examinations and drugs into their corresponding categories. To ease the manual inspection of the result, a worthwhile subset of rules, i.e., the non-redundant generalized rules, is considered. Furthermore, rules are classified according to the involved data features (medical treatments or patient profiles) and then explored in a top-down fashion, i.e., from the small subset of high-level rules a drill-down is performed to target more specific rules.

The experiments, performed on a real diabetic patient dataset, demonstrate the effectiveness of the proposed approach in discovering interesting rule groups at different abstraction levels.
Keywords: Healthcare Informatics, Data Mining, Generalized Association Rule Mining

---

1. Introduction

Healthcare systems are nowadays integrated platforms that can take advantage of advanced data management and analysis solutions. Large amount of data on medical patient history is commonly stored by healthcare organizations. Data mining techniques can be used to analyze these large data collections and to extract knowledge useful for physicians and healthcare organizations.

This study addresses the problem of analyzing patient historical data to identify valuable correlations among patient treatments and profiles. The extracted patterns allow experts to (i) identify the medical treatments com-

monly followed by patients with a given disease, (ii) verify the adherence of medical treatments to shared medical guidelines, (iii) improve the effectiveness of medical treatments, and (iv) plan resource allocation and reduce costs incurred by organizations.

Association rule extraction is an established data mining technique to discover interesting correlations among large datasets [44]. Since patient history data is relatively sparse, discovering association rules from these datasets is a challenging task. Discovering correlations among data items that rarely co-occur may become computationally intractable when coping with large datasets. On the other hand, focusing only on most frequent item recurrences could provide not fruitful enough information. Furthermore, since a large rule set could be mined, inferring useful and actionable knowledge from the extracted rules can be a complex task.

This paper presents: (i) MeTA (Medical Treatment Analysis), a new data analysis framework targeted to the discovery of underlying multiple-level correlations among patient treatments and profiles. (ii) The classification of the mined rules into classes according to the represented data features (e.g., examinations, drugs). (iii) The exploration of rules in order of descending level of abstraction of the represented information in the input taxonomy. (iv) The application of MeTA to a real-life use case, i.e., the analysis of diabetic patient data provided by the National Health Center (NHC) of an Italian province.

Patients datasets consist of log files holding information about patient treatments and census data. Each row contains a set of pairs (*feature*, *value*), where *feature* corresponds to a specific data feature (i.e., *Examination*, *Drug*,

Age, or *Gender*), while *value* is the corresponding feature value. MeTA discovers interesting and multiple-level correlations among patient data called generalized rules [39]. These rules are represented in the form $X \rightarrow Y$, where X and Y are disjoint sets of items (called itemsets). The implication means that (i) itemsets X and Y frequently co-occur in the analyzed dataset (regardless of the temporal order of occurrence of X and Y in the source data), (ii) the strength of the implication between X and Y is above a given threshold, and (iii) X and Y may also include items belonging to different abstraction levels. Item generalization is driven by a taxonomy, which consists of a set of is-a hierarchies built over the analyzed data. For example, drugs can be generalized based on the addressed pathology [7], while examinations are clustered based on the focused area (e.g., liver or cardiovascular system). Aggregating items into higher-level concepts (e.g., examinations into the corresponding category) prevents the discarding of potentially useful knowledge and thus counteracts the issue of data sparseness. In our context of analysis, we disregard the temporal order of prescriptions and we specifically focus on discovering multiple-level co-occurrences among examination/drug prescriptions. In Section 3 we will demonstrate that these patterns are worth considering for targeted analysis (e.g., resource allocation, healthcare service management). To make the mined result more manageable by domain experts for manual inspection, MeTA considers a worthwhile rule subset, i.e., the non-redundant rules [48]. Non-redundant rules are generated from closed itemsets [33], which are a compact and non-redundant subset of frequent itemsets. Furthermore, MeTA categorizes the rules into four groups according to the represented data facet (e.g., examination, drugs).

Within each group rules are further classified according to their level of abstraction of the contained items in the input taxonomy. The usefulness of both non-redundant rule selection and rule categorization for improving the manageability of the mining result is discussed in Section 3.4.

As an example, let us consider rule {(*Examination*, *Routine*), (*Examination*, *Cardiovascular*)} $\rightarrow$ {(*Drug*, *Blood and blood forming organs Category*)}. It indicates that drugs in category "Blood and blood forming organs" have been prescribed to a relatively large number of patients to whom routine and cardiovascular examinations have been prescribed as well (disregarding the temporal order of prescriptions). This information could be deemed to be useful, for example, for shaping drug provision to medical divisions according to the most commonly performed examinations. The rule is high-level, because it contains only examination and drug categories. Conversely, rules containing *also* or *only* single examinations/drugs will be denoted as cross- or low-level rules, respectively. The cross-level rule {(*Examination*, *Routine*), (*Examination*, *Cardiovascular*)} $\rightarrow$ {(*Drug*, *Acetylsalicylic Acid*)} can be extracted as well in case drug Acetylsalicylic Acid has predominantly been prescribed among the "Blood and blood forming organs" drug category. Note that the aforesaid high- and cross-level rules are more likely to be frequent than their low-level descendant rules (e.g., {(*Examination*, *Complete blood count*), (*Examination*, *Cholesterol*)} $\rightarrow$ {(*Drug*, *Acetylsalicylic Acid*)}). High- and cross-level rules are worth considering because (i) they represent, from a high-level viewpoint, valuable information that may remain hidden in sparse datasets at lower abstraction levels and (ii) they are typically more manageable than low-level rules for manual result exploration.

We assessed the usability of MeTA on a real dataset of diabetic patients provided by the National Health Center (NHC) of an Italian province. The experiments demonstrate that, starting from a large collection of raw data on patient history, the framework allows experts to identify several interesting high-, cross-, and low-level correlations among patient treatments and profiles. The results were validated by clinical domain experts. The extracted rules appear to be consistent with the guidelines for diabetes disease [1, 23, 24]. Furthermore, the extraction of high- and cross-level rules appears to effectively overcome limitations of traditional approaches.

This paper is organized as follows. Section 2 presents the architecture of the proposed framework and it describes its main blocks. Section 3 assesses the effectiveness of the system in performing knowledge discovery from a real diabetic patient dataset. Section 4 compares our approach with most relevant related works, while Section 5 draws conclusions and presents future developments of this work.

2. The Medical Data Generalized Rule Miner system

MeTA (Medical Treatment Analysis) is a novel framework for medical data analysis, which focuses on characterizing medical treatments at different granularity levels.

The main MeTA architectural blocks are depicted in Figure 1. A brief description of each block follows.

Data collection and preparation. This block aims at making information about patient characteristics, examinations, and drugs suitable for the mining

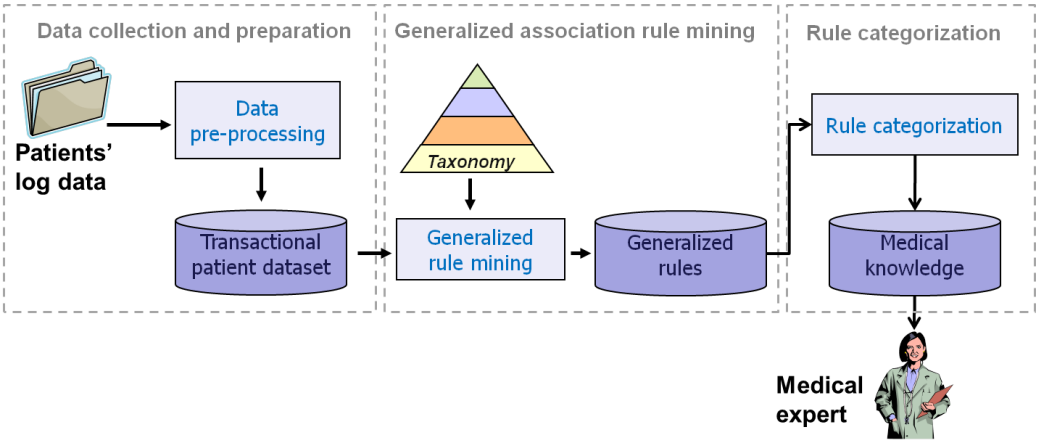

Figure 1: The Medical Treatment Analysis framework

process. Patient datasets are tailored to a transactional data format, where each transaction corresponds to a different patient and it consists of a set of items, which represent patient census data (e.g., age, gender), prescribed examinations (e.g., Glucose level), or prescribed drugs (e.g., Acetylsalicylic Acid). Transactional datasets are enriched with an (analyst-provided) taxonomy built over the data items.

Generalized association rule mining. This block focuses on discovering multiple-level correlations among the preprocessed data in the form of generalized association rules. The extraction process is driven by the input taxonomy to generalize data items at higher abstraction levels. To extract only the rules that (i) occur frequently and (ii) represent positively correlated implications among pairs of item sets in the source dataset, rules are filtered according established quality measures, i.e., support and lift [43]. Furthermore, to filter out less informative rules only the subset of non-redundant

rules [48] is considered for further analyses.

Rule categorization. To make the mining result manageable by experts for manual inspection, rules are categorized according to their represented information. To analyze correlations among patient data regardless of the patient profile, rule subsets that represent (i) correlations among examinations, (ii) correlations among drugs, and (iii) correlations between examinations and drugs are analyzed separately. On the other hand, to gain more insights into specific user profiles (e.g., elderly men, kids) implications between specific patient characteristics and examinations/drugs are considered. To easily explore rule categories the corresponding rules are further classified as high-level, cross-, or low-level according to the level of abstraction of the contained information in the input taxonomy.

A more thorough description of each block is reported below.

2.1. Data collection and preparation

Healthcare systems usually collect heterogeneous personal information about patients into log datasets. For example, the list of prescribed examinations and drugs is stored in separate log files to allow doctors to keep track of diagnosis and therapies and healthcare system managers to plan purchases and resource allocations. In parallel, to characterize the patient population, census data about patients, such as gender and age, are usually collected in separate datasets.

The MeTA framework collects and stores into a unique data repository these three main patient data types. More specifically, for each patient the

list of (i) prescribed examinations, (ii) drugs, and (iii) the main patient characteristics are stored.

To enable the mining process, the collected patient data is tailored to a transactional data format. A transactional patient dataset is a set of transactions, where each transaction corresponds to a patient and it consists of a set of patient features, called items. Items can be related to examinations (e.g., *Glucose level*), drugs (e.g., *Acetylsalicylic Acid*), or patient census data (e.g., *Male*). In this work we focus on age and gender as peculiar patient census data. Items are represented in the form (*feature*,*value*), where *feature* is *Examination*, *Drug*, *Age*, or *Gender*, while *value* is the corresponding feature value. A more formal definition of transactional patient dataset is given below.

Definition 1. Transactional patient dataset. *Let E be the set of all possible patient examinations, M the set of all possible drugs, and C the set of census data features. Let $\Omega(c_i)$ be the domain of an arbitrary census data feature $c_i \in C$ (i.e., the set of all possible values assumed by c_i). An item is a pair (feature,v_i), where $v_i \in E$ if feature=Examination, $v_i \in M$ if feature=Drug, and $v_i \in \Omega(c_i)$ if feature=c_i. A transactional patient dataset $\boldsymbol{D}$ is a set of transactions, where each transaction $t_i \in \boldsymbol{D}$ is a set of items.*

Let us consider, as running example, the dataset reported in Table 1. It consists of 5 records, each one related to a different patient. For each patient the identifier (Pid), age (Age), gender (Gender), and a list of prescribed examinations and drugs is given. The dataset contains four different examinations (*HDL Cholesterol*, *Glucose level*, *Electrocardiogram*, and *Blood*

Table 1: Example of patient transactional dataset.

Pid	Transaction
1	{(*Age*,*Elder*), (*Gender*,*Male*), (*Examination*,*HDL Cholesterol*), (*Examination*,*Glucose level*), (*Examination*,*Electrocardiogram*), (*Drug*,*Acetylsalicylic Acid*)}
2	{(*Age*,*Elder*), (*Gender*,*Female*), (*Examination*,*Glucose level*), (*Drug*,*Acetylsalicylic Acid*), (*Drug*,*Moxifloxacin*)
3	{(*Age*,*Elder*), (*Gender*,*Male*), (*Examination*,*Glucose level*), (*Examination*, *Electrocardiogram*), (*Examination*,*Blood count*), (*Drug*,*Moxifloxacin*)
4	{(*Age*,*Teenager*), (*Gender*,*Female*), (*Examination*,*Glucose level*), (*Drug*,*Acetylsalicylic Acid*), (*Drug*,*Moxifloxacin*)
5	{(*Age*,*Elder*), (*Gender*,*Male*), (*Examination*,*Electrocardiogram*), (*Examination*,*Blood count*), (*Drug*,*Acetylsalicylic Acid*)

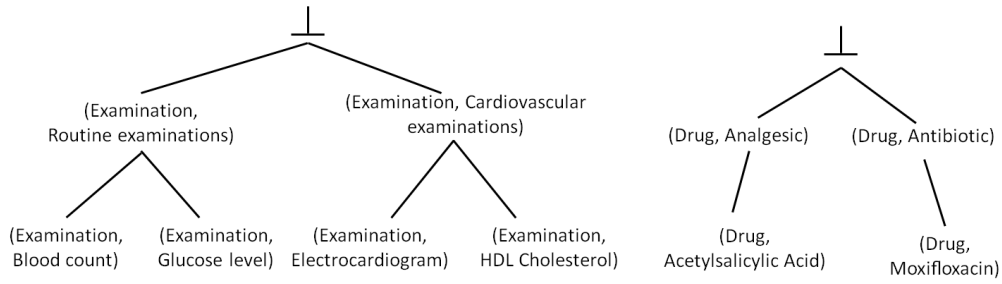

Figure 2: Example of taxonomy built over the transactional patient dataset.

count) and two different drugs (*Acetylsalicylic Acid* and *Moxifloxacin*). For example, patient with Pid 5 is an elderly man to whom examinations *Electrocardiogram* and *Blood count* have been prescribed at least once. Furthermore, he has already taken *Acetylsalicylic Acid* but not *Moxifloxacin*.

To enable the process of generalized rule mining from a transactional patient dataset **D**, a taxonomy (i.e., a set of generalization hierarchies) is built over the items in **D**. The taxonomy aggregates examinations and drugs into high-level concepts, i.e., examinations are generalized as examination categories while drugs as drug categories.

Definition 2. Taxonomy. *Let* $\boldsymbol{D}$ *be a transactional patient dataset and* $\boldsymbol{I}$ *the set of items in* $\boldsymbol{D}$*. A generalization hierarchy* GH_{I_k} *(*$I_k \subseteq \boldsymbol{I}$*) built over*

D *is a predefined hierarchy of aggregations defined over a subset of items in* ***I****, where hierarchy leaves are items in* ***I****, while non-leaf nodes in* GH_{I_k} *are ancestors of their corresponding children. Each hierarchy has a root node (denoted as* $\perp$*) which aggregates all its items. A taxonomy* ***T*** *built over* ***D*** *consists of a set of generalization hierarchies* GH_{I_k} *for which* $\cup_{GH_{I_k} \in \boldsymbol{T}} I_k = \boldsymbol{I}$*.*

Although taxonomies can potentially contain many generalizations over the same item (e.g., many categories for the same examination), in this work we will consider only taxonomies containing at most one generalization per item.

An example of taxonomy built over the running example dataset is reported in Figure 2. Examinations *Blood count* and *Glucose level* are classified as *Routine examinations*, whereas examinations *Electrocardiogram* and *HDL Cholesterol* are generalized as *Cardiovascular*. Finally, drugs *Acetylsalicylic Acid* and *Moxifloxacin* are classified as *Analgesic* and *Antibiotic*, respectively.

2.2. Generalized association rule mining

This block focuses on discovering multiple-level associations, in the form of generalized association rules, from the transactional patient dataset **D** with a taxonomy **T**.

Association rules represent underlying correlations among the analyzed data items [2]. More specifically, an association rule is an implication $A \Rightarrow B$, where A and B are itemsets, i.e., sets of data items. A k-itemset I a set of items of size k that occurs in **D**.

For example, {(*Examination,Glucose level*),(*Examination,Electrocardiogram*)} is a 2-itemset that represents the co-occurrence of two specific examinations

in medical treatments, while the association rule {(*Examination,Glucose level*)} $\rightarrow$ {(*Examination,Electrocardiogram*)} indicates that the occurrence of examination *Glucose level* "implies" those of examination *Electrocardiogram* in the analyzed data.

Generalized association rules [39] are rules that may also contain items at higher abstraction levels, i.e., the generalized items. Every item that is associated with a non-leaf node of the taxonomy **T** (see Definition 2) is considered as a generalized item. Similarly, generalized itemsets are itemsets including at least one generalized item.

Definition 3. Generalized itemset. *Let* $\boldsymbol{D}$ *be a transactional patient dataset and* $\boldsymbol{I}$ *be the set of distinct items in* $\boldsymbol{D}$*. Let* $\boldsymbol{T}$ *be a taxonomy built over* $\boldsymbol{D}$ *and* $\boldsymbol{G}$ *the set of generalized items (high-level tag aggregations) derived by all the generalization hierarchies in* $\boldsymbol{T}$*. A generalized itemset I is a subset of* $\boldsymbol{I} \bigcup \boldsymbol{G}$ *including at least one generalized item in* $\boldsymbol{G}$*.*

For example, according to the taxonomy in Table 2, {(*Examination,Routine*),(*Examination,Elect*
is a generalized itemset.

Generalized itemsets are characterized by two quality indexes, i.e., the level and support. The level of a generalized item i_k with respect to a taxonomy indicates the degree of abstraction of the represented information.

Definition 4. Generalized itemset level. *Let* $\boldsymbol{D}$ *be a transactional patient dataset and* $\boldsymbol{I}$ *be the set of distinct items in* $\boldsymbol{D}$*. Let* $\boldsymbol{T}$ *be the taxonomy defined over* $\boldsymbol{D}$ *and* i_k *an arbitrary item or generalized item in* $\boldsymbol{T}$*. The level of (generalized) item* i_k *is defined as the height of* $\boldsymbol{T}$*'s subtree rooted in* i_k*.*

The level of a generalized itemset is defined as the maximum level among the levels of its items.

Generalized itemsets whose items have all the same level are called *level-sharing itemsets* [22]. The level of a level-sharing itemset with respect to a taxonomy corresponds to the one of its items.

The support of a generalized itemset evaluates its observed frequency of occurrence in the analyzed data. It is defined in terms of the itemset coverage with respect to the analyzed data.

Definition 5. Generalized itemset coverage. *Let $\boldsymbol{D}$ be a transactional patient dataset and $\boldsymbol{T}$ the corresponding taxonomy. A (generalized) itemset I covers a given transaction $t_i \in \boldsymbol{D}$ if all its (possibly generalized) items $i_k \in I$ are either included in t_i or ancestors (generalizations) of items $i_k \in t_i$ with respect to $\boldsymbol{T}$.*

The support of a generalized itemset I is given by the ratio between the number of transactions $t_i \in \mathbf{D}$ covered by I and the cardinality of $\mathbf{D}$.

A (generalized) itemset I is said to be a descendant of another generalized itemset Y if (i) I and Y have the same length (i.e., the same number of items) and (ii) for each item $y \in Y$ there exists an item $i \in I$ that is a descendant of y.

The concept of generalized association rule extends traditional association rules to the case in which they may include either generalized or not generalized itemsets. A more formal definition follows.

Definition 6. Generalized association rule. *Let A and B be two (generalized) itemsets. A generalized association rule is represented in the form*

$R : A \Rightarrow B$, *where* A *and* B *are the body and the head of the rule respectively.*

A and B are also denoted as antecedent and consequent of the generalized rule $A \Rightarrow B$. Generalized association rule extraction is commonly driven by rule support (s) and confidence (c) quality indexes [39]. While the support index represents the observed frequency of occurrence of the rule in the source dataset, the confidence index represents the rule strength.

Definition 7. Generalized association rule support. *Let* $\boldsymbol{D}$ *be a transactional patient dataset and* $\boldsymbol{T}$ *a taxonomy. The support of a generalized rule* $R : A \Rightarrow B$ *is defined as the support (i.e., the observed frequency) of* $A \cup B$ *in* $\boldsymbol{D}$*.*

Definition 8. Generalized association rule confidence. *Let* $\boldsymbol{D}$ *be a transactional patient dataset and* $\boldsymbol{T}$ *a taxonomy. The confidence of a rule* $R : A \Rightarrow B$ *is the conditional probability of occurrence in* $\boldsymbol{D}$ *of the generalized itemset* B *given the generalized itemset* A*.*

For example, the generalized association rule {(*Examination,Routine*)} $\rightarrow$ {(*Examination,Electrocardiogram*)} (s=60%,c=100%) indicates that examinations belonging to category *Routine* co-occur with examination *Electrocardiogram* in $\frac{3}{5}$ of the transactions of the analyzed dataset (Pids 1, 3, and 5) and the implication holds in $\frac{3}{3}$=100% of the cases.

In some cases, measuring the strength of a rule in terms of support and confidence may be misleading [42]. When the rule consequent is characterized by relatively high support value, the corresponding rule may be characterized by a high confidence even if its actual strength is relatively low. To overcome

this issue, the lift (or correlation) index [43] may be used, rather than/beyond the confidence index, to measure the (symmetric) correlation between body and head of the extracted rules.

Definition 9. Generalized association rule lift. *Let $A \rightarrow B$ be an association rule. Its lift is given by*

$$\mathrm{l(A,B)} = \frac{\mathrm{c}(A \rightarrow B)}{\mathrm{s(B)}} = \frac{\mathrm{s}(A \rightarrow B)}{\mathrm{s(A)s(B)}} \tag{1}$$

where $\mathrm{s}(A \rightarrow B)$ *and* $\mathrm{c}(A \rightarrow B)$ *are, respectively, the rule support and confidence, and* $\mathrm{s}(A)$ *and* $\mathrm{s}(B)$ *are the supports of the rule antecedent and consequent.*

If l(A,B) is equal to or close to 1, itemsets A and B are not correlated with each other. Lift values significantly below 1 show negative correlation, whereas values significantly above 1 indicate a positive correlation between itemsets A and B, i.e., the implication between A and B holds more than expected. For example, rule $\{(\textit{Examination,Routine})\} \rightarrow \{(\textit{Examination,Electrocardiogram})\}$ is positively correlated, because its lift value is equal to $\frac{5}{3}$.

Since the interest of uncorrelated or negatively correlated rules is marginal in our context of analysis, MeTA only considers frequent and positively correlated generalized association rules. Specifically, given a transactional patient dataset, a taxonomy, a minimum support threshold (*minsup*), and a minimum lift threshold (*minlift*) MeTA discovers all the generalized association rules whose:

- support value is above a given minimum support threshold *minsup*, i.e., s(R)>*minsup*, and

- lift value is above a given minimum lift threshold *minlift*, i.e., l(R)>*minlift*.

The generalized rules that satisfy all the above conditions will be denoted as *strong rules* throughout the paper.

Since the set of strong rules may still contain less informative rules, a further pruning step is applied prior to performing further analyses. Specifically, MeTA discovers non-redundant generalized rules [48], which are a worthwhile subset of strong generalized rules. Extensions of a strong generalized rule are classified as redundant if they have the same support and confidence of their specialized version. A more formal definition is given below. It extends the concept of non-redundant rule, first proposed in [48], to the case in which rules may also contain generalized items.

Definition 10. Non-redundant generalized association rule. *Let* $R : A \Rightarrow B$ *be a strong generalized rule.* R *is non-redundant if there exists no strong rule* $R^* : C \Rightarrow D$, $C \subseteq A \wedge D \subseteq B$ *such that the support and confidence of* R *and* R^* *are equal.*

To generate non-redundant generalized rules we used the publicly available implementation of the algorithm proposed in [48] on an extended dataset version, in which transactions contain both items and their corresponding generalizations according to the input taxonomy. This generalized rule mining strategy is similar to the one previously adopted in [39] in the context of market basket analysis.

2.3. Rule categorization

The generalized rules extracted during the last MeTA step are explored by domain experts to discover valuable information. Unfortunately, when

coping with relatively large or complex transactional patient datasets the number of mined rules could be so large that a manual inspection becomes unfeasible. To overcome this issue, this block focuses on categorizing the extracted rules into homogeneous groups, according to their represented information.

MeTA partitions rules into worthwhile subsets that characterize the underlying data from different viewpoints, because they contain different combinations of patient features and/or medical treatments. We highlighted four representative rule classes, which are thoroughly described below. Table 2 reports the rule template for each class.

Class E-Rules: Correlations between examinations. Rules in this group represent correlations among examinations regardless of the characteristics of the analyzed patients and prescribed drugs. For example, $\{(Examination, Routine)\} \rightarrow \{(Examination, Electrocardiogram)\}$ belongs to Class E-Rules. This class may potentially include more complex rules, such as $\{(Examination, Routine), (Examination, Blood\ count)\} \rightarrow \{(Examination, Electrocardiogram)\}$. In other words, rule antecedent can be, in general, itemsets of arbitrary size.

Class D-Rules: Correlations between drugs. Rules in this group focus the experts' attention on correlations among the prescribed drugs, disregarding examinations and patient characteristics. For example, $\{(Drug, Acetylsalicylic\ Acid)\} \rightarrow \{(Drug, Moxifloxacin)\}$ belongs to Class D-Rules. Even in this class rules can represent implications where the antecedent is an itemset of arbitrary size.

Table 2: Rule categories. $*^1$ represents an examination or an examination class, $*^2$ represents either a drug or a drug class $*^3$ represents either an age or an age group, while $*^4$ is a gender value (male or female).

Class ID	**Name**	**Template**
E-Rules	*Correlations between examinations*	$\{(Examination,*^1)\} \rightarrow \{(Examination,*^1)\}$
D-Rules	*Correlations between drugs*	$\{(Drug,*^2)\} \rightarrow \{(Drug,*^2)\}$
ED-Rules	*Correlations between examinations and drugs*	$\{(Drug,*^2)\} \rightarrow \{(Examination,*^1)\}$ $\{(Examination,*^1)\} \rightarrow \{(Drug,*^2)\}$ $\{(Examination,*^1),(Drug,*^2)\} \rightarrow \{(Examination,*^1)\}$ $\{(Examination,*^1),(Drug,*^2)\} \rightarrow \{(Drug,*^2)\}$
P-Rules	*Profile-based correlations*	**Age Profiles** $\{(Age,*^3)\} \rightarrow \{(Examination,*^1)\}$ $\{(Age,*^3)\} \rightarrow \{(Drug,*^2)\}$ $\{(Age,*^3),(Examination,*^1)\} \rightarrow \{(Drug,*^2)\}$ $\{(Age,*^3),(Drug,*^2)\} \rightarrow \{(Examination,*^1)\}$ **Gender Profiles** $\{(Gender,*^4)\} \rightarrow \{(Examination,*^1)\}$ $\{(Gender,*^4)\} \rightarrow \{(Drug,*^2)\}$ $\{(Gender,*^4),(Examination,*^1)\} \rightarrow \{(Drug,*^2)\}$ $\{(Gender,*^4),(Drug,*^2)\} \rightarrow \{(Examination,*^1)\}$ **Age-Gender Profiles** $\{(Age,*^3),(Gender,*^4)\} \rightarrow \{(Examination,*^1)\}$ $\{(Age,*^3),(Gender,*^4)\} \rightarrow \{(Drug,*^2)\}$ $\{(Age,*^3),(Gender,*^4)\} \rightarrow \{(Examination,*^1)\}$ $\{(Age,*^3),(Gender,*^4)\} \rightarrow \{(Drug,*^2)\}$

Class ED-Rules: Correlations between examinations and drugs. This group of rules represents co-occurrences between drugs and examinations into the patient dataset, regardless of patient characteristics. More specifically, all the rules that contain both examinations/examination category and drugs/drug categories into their antecedent/consequent are assigned to class ED-Rules. For example, {(*Examination,Routine*), (*Drug,Aspirin*)} $\rightarrow$ {(*Examination,Electrocardiogram*)} is assigned to this class. It indicates the association between the co-occurrence of an examination category and a drug and a specific examination.

Class P-Rules: Profile-based correlations. The former rule classifications disregard patient characteristics. Nevertheless, experts can deem such information to be useful for characterizing specific user profiles (e.g., elderly men, kids). This class consists of all the rules that contain any item related to a census feature in their rule antecedent. This rule subset can be further categorized according to the considered census features, because each combination of patient census features may represent a distinct and potentially meaningful user profile. Since in our work we target our analysis on age and gender census features, rules belonging to Class P-Rules can be partitioned into the three subgroups reported in Table 2.

For example, {(*Age,Elder*), (*Gender,Male*)} $\rightarrow$ {(*Examination,HDL Cholesterol*)} indicates that the *HDL Cholesterol* examination has frequently been prescribed to elderly men. Similarly, rule {(*Age,Elder*), (*Drug,Acetylsalicylic Acid*)} $\rightarrow$ {(*Examination,HDL Cholesterol*)} indicates that the *HDL Cholesterol* examination has frequently been prescribed to elderly people (males or

females) who have taken drug Acetylsalicylic Acid (regardless of the temporal order of drug/examination prescriptions).

2.3.1. Level-wise exploration of rule categories

Given a worthy set of rule categories, experts are asked to go into detail about the contained rules. However, since generalized rules potentially represent information at different levels of granularity, rule class exploration could be challenging unless considering taxonomy abstraction levels as reference information.

To easily explore rule categories, the corresponding rules are further classified as high-, cross-, or low-level according to the level of abstraction of the contained information in the input taxonomy.

High-level rules are generalized rules $A \rightarrow B$, where A and B are level-sharing itemsets with the same level $l > 1$. They typically represent general knowledge and thus they should be considered first during manual result exploration.

For example, $\{(Examination, Routine)\} \rightarrow \{(Examination, Cardiovascular\ examination)\}$ is a high-level rule, because both rule antecedent and consequent are level-2 itemsets.

Cross-level rules are generalized rules $A \rightarrow B$, where A and B are either not level-sharing itemsets or level-sharing itemsets with different level. They combine detailed and general information by climbing up and down the taxonomy for different data features. Given a subset of high-level rules, cross-level rules can be considered as an intermediate step to perform drill-down (i.e., moving from general to detailed information).

For example, $\{(Examination, Blood\ count)\} \rightarrow \{(Examination, Cardiovascular$

examination)} is a cross-level rule, because the rule antecedent is a level-1 itemset, whereas the rule consequent is a level-2 itemset. If the high-level rule {(*Examination,Routine*)} $\rightarrow$ {(*Examination,Cardiovascular examination*)} is deemed to be useful for advanced analysis, then considering the former cross-level rule can be relevant to analyze the underlying correlations between a specific routine examination and the Cardiovascular examination category.

Low-level rules are not generalized rules $A \rightarrow B$, i.e., both A and B are not generalized (level-1) itemsets. They typically represent very detailed knowledge. When coping with relatively sparse datasets, many of these rules could be discarded during the mining process by enforcing the minimum support threshold. However, their peculiar information is likely to be covered, to a certain extent, by cross- and high-level rules. For example, {(*Examination,Urine test*)} $\rightarrow$ {(*Examination,Electrocardiogram*)} is an example of low-level rules. These rules can be analyzed to gain more insights on a specific subset of cross-level or high-level rules.

3. Experimental results

We performed various experiments on a real-life dataset collected by an Italian Health Center to demonstrate effectiveness and efficiency of the MeTA framework.

The experimental section is organized as follows. Section 3.1 describes the main characteristics of the real-life dataset analyzed in this study, while Section 3.2 summarizes the most relevant results achieved during the mining session as well as highlights the significance and usability of the discovered

high-level correlations among data. Section 3.3 analyzes the distribution of the discovered rules across categories and levels of abstraction of the extracted knowledge. A quantitative analysis of the complexity of the rule exploration process is given in Section 3.4. Finally, Section 3.5 analyzes the efficiency of MeTA in terms of execution time.

All the experiments were performed on a quad-core 3.30 GHz Intel Xeon workstation with 16 GB of RAM, running Ubuntu Linux 12.04 LTS. The software used to perform rule extraction and post-processing is available online at [30].

3.1. Diabetic patient dataset and taxonomy

The dataset considered in this study was collected by an Italian Local Health Center (LHC). Specifically, in 2007 they collected into a unique LHC dataset all the accesses to the medical center year-round. Then, from the LHC dataset the examination log data of all the patients with overt diabetes were extracted. Raw data consist of 95,788 records and they include examinations and drugs prescribed to 3,565 patients. The dataset contains information about male and female patients in a wide age range (i.e., between 4 and 95 years). To analyze diabetes complications at various degrees of severity both routine and more specific examinations were recorded jointly with prescribed drugs. The diagnostic and therapeutic procedures were defined using the ICD 9-CM (International Classification of Diseases, 9th revision, Clinical Modification) [23]. Drugs were identified by the pharmaceutical coding system adopted by the Anatomical Therapeutic Chemical (ATC) Classification System [7].

The generalization hierarchy over examinations is shown in Table 3. It

contains 26 examinations clustered into 7 examination categories. The selected examination categories are based on the expert-driven classification reported in [5].

The drug generalization hierarchy contains as leaves the drugs encoded by using the fifth level of the ATC classification system defined in [7]. Drugs are aggregated into the corresponding drug category, according to the first level of the standard ATC classification system. For instance, drug acetylsalicylic acid (i.e., code: B01AC06) is a leaf node of the drug generalization hierarchy and Category B (i.e., Category Blood and blood forming organs) is its generalization. Our dataset contains 200 distinct drugs and 14 distinct categories. Table 4 reports the hierarchy defined over drugs.

Human life is often divided into various age ranges (e.g., infancy, middle-adulthood, old age). Age feature values have been discretized into the following 8 age groups, which represent established ranges of the human lifespan [46]: [0-6], [7-12], [13-22], [23-39], [40-59],[60-75], [76-90], and [91-101].

3.2. Analysis of the mined rules

We performed several generalized rule extractions from the patient dataset by enforcing different minimum support (*minsup*) and lift (*minlift*) thresholds. To perform knowledge discovery from the mined rules, we selected a representative configuration setting, i.e., we set *minsup* to 1% and *minlift* to 1.1. The reasons behind the choice of the support threshold are twofold. Firstly, too low/high support threshold values yield very detailed/general rule sets and thus they may not produce manageable yet interesting knowledge. Secondly, it is well-known that averagely low-support rules commonly represent potentially interesting knowledge if they represent positive corre-

Table 3: Generalization hierarchy over examinations.

Examination category	**Examination**
Routine examinations	*Checkup visit*
	Glucose level
	Urine test
	Venous blood
	Complete blood count
	Hemoglobin
Cardiovascular examinations	*Electrocardiogram*
	Cholesterol
	HDL Cholesterol
	Triglycerides
Eye examinations	*Fundus oculi*
	Angioscopy
	Complete eye examination
	Retinal photocoagulation
Liver examinations	*AST*
	ALT
	Bilirubin
	Gamma GT
Kidney examinations	*Urin acid*
	Microscopic urine analysis
	Culture urine
	Creatinine clearance
	Creatinine
	Microalbuminuria
Carotid examinations	*ECO Doppler carotid*
Limb examinations	*ECO Doppler limb*

Table 4: Generalization hierarchy over drugs.

Drug category	**Drug**
Category A: Alimentary tract and metabolism	*A01AA01: Sodium fluoride* *A05AX01: Piprozolin* *...*
Category B: Blood and blood forming organs	*B01AC06: Acetylsalicylic acid* *B03AA03: Ferrous gluconate* *...*
Category C: Cardiovascular system	*C09AA05: Ramipril* *C10AA07: Rosuvastatin* *...*
Category D: Dermatologicals	*D01AA02: Natamycin* *D01AA03: Hachimycin* *...*
Category G: Genito-urinary system and sex hormones	*G04CB01: Finasteride* *G04CX03: Mepartricin* *...*
Category H: Systemic hormonal preparations, excluding sex hormones and insulins	*H02AA02: Fludrocortisone* *H02AB07: Prednisone* *...*
Category J: Antiinfectives for systemic use	*J01MA12: Levofloxacin* *J02AC04: Posaconazole* *...*
Category L: Antineoplastic and immunomodulating agents	*L01AA07: Trofosfamide* *L01AB01: Busulfan* *...*
Category M: Musculo-skeletal system	*M03AC10: Mivacurium chloride* *M03BA05: Febarbamate* *...*
Category N: Nervous system	*N04AA02: Biperiden* *N04AB01 Etanautine* *...*
Category P: Antiparasitic products, insecticides and repellents	*P01AA04: Chlorquinaldol* *P01AC01: Diloxanide* *...*
Category R: Respiratory system	*R03AC02: Salbutamol* *R03BA02: Budesonide* *...*
Category S: Sensory organs	*S02AA10: Acetic acid* *S02BA03: Prednisolone* *...*
Category V: Various	*V10XX01: Sodium phosphate* *V10XA01: Sodium iodide* *...*

lations among items (i.e., their lift is above 1) [15]. Nevertheless, by generalizing items at higher abstraction levels some of the low-support correlations among data are still represented by higher-level rules. Hence, we selected *minsup*=1% as a good trade-off between rule set specialization and generality. We also enforced a minimum lift threshold equal to 1.1 to prune both negatively correlated and uncorrelated item combinations. On the one hand, the interest of negatively correlated rules (i.e., rules with lift below 1) is marginal in our context of analysis. On the other hand, rules with lift close to 1 are misleading because their occurrences are not actually correlated with each other. Hence, among the positively correlated rules we further pruned approximately 10% of them whose lift value is between 1 and 1.1. Finally, we focused on the rules with length below or equal to 3, i.e., the rules consisting of pairs or triples of (generalized) items, because they represent the most actionable correlations among drugs/examinations. However, to specialize the rules that provide peculiar information, experts could performed further extractions and explore longer rules complying with the given category.

We first analyzed the rules than hold for all patients, i.e., we considered rules belonging to Classes E-Rules, D-Rules, and ED-Rules (Section 2.3). Then, we focused on rules that concern patient profiles (i.e., Class P-Rules).

3.2.1. Correlations between examinations and drugs

Tables 5 and 6 report worthwhile subsets of correlations between sets of examinations and drugs, respectively. The former rules represent potentially interesting correlations among examinations, whereas the latter correlations among drugs. A worthy subset of correlations between examinations and drugs (Class ED-Rules) is summarized in Table 7. For each rule, we reported

support (percentage), confidence (percentage), lift, and the corresponding type, according to the level-dependent classification reported in Section 2.3.1 (low-level, cross-level, high-level).

Analysis of correlations between examinations (Class E-Rules). This section addresses the analysis of a subset of interesting correlations between examinations. First, we are particularly interested in analyzing the co-occurrences among examination categories, while disregarding the temporal order of prescriptions. High-level rules, such as rules (1)-(7) in Table 5, represent positive correlations between examination categories. They can be used to target the analysis towards specific issues. For example, rules (1) and (2) highlight a pairwise association between liver and kidney examinations, which hold 2.52 times more than expected according to the corresponding lift value[1]. In other words, the expected frequency of co-occurrence of the two examination category (assuming the independence between the occurrences of the single examination categories) is significantly lower than the observed one. The high-level rules (1) and (2) can be used to efficiently schedule medical examination timetables according to their corresponding prescriptions. For example, since liver and kidney examinations are frequently prescribed to the same patient, scheduling both examinations at the same day could reduce patient recovery time. A deeper insight into liver and kidney examinations may be focused on (i) assessing the adherence of medical treatments to the medical guidelines suggested by the Italian Ministry of Health about liver and kidney diseases in diabetic patients or (ii) proposing new guidelines

[1]The lift value of the two rules is the same because of the symmetry of the lift measure [43].

according to the observed correlations between specific liver and kidney examinations. Similar analyses can be performed starting from the pairwise correlations between the examination categories represented by rules (3)-(6). Since association rules can also represent higher-order associations among data, we should not restrict our analyses to pairwise associations among items. For example, rule (7) shows a positive correlation between liver, cardiovascular, and kidney examination categories. Longer rules can be used either to specialize known lower-order associations or to figure out new and more complex medical treatments.

To deepen into the analysis of the most specific correlations between examinations, high-level rules are often not enough. In fact, they provide a high-level view of the underlying correlations among data, which could be insufficient to perform targeted analysis. On the other hand, as discussed in the following, high-level rules are very important because they also represent those patterns that have not been separately extracted at lower abstraction levels because they are infrequent according to the support threshold.

A step forward is to consider also cross-level rules, which contain both low- and high-level information, i.e., examinations and examination categories, at the same time. To take advantage of the preliminary analysis of high-level rules, only the subset of cross-level rules that are related to some interesting high-level rule are considered. For example, based on rule (1), we can deepen our analysis into the search of underlying correlations between specific examinations and examination categories. For instance, given the subset of patients to whom liver examinations are frequently prescribed, what specific kidney examination is most likely to be frequently prescribed

as well? From the comparison between the confidences of rules (8)-(13), uric acid appears to be the most likely kidney examination, because to 74.8% of the patients associated with a liver examination the uric acid examination has been prescribed as well. This information is worthy because it gives more insights into a subset of medical treatments. Similarly, other combinations of examinations and examination categories (which have been omitted for the sake of brevity) have been mined.

The last step is the analysis of low-level rules, which represent significant correlations among single examinations (disregarding the examination categories). The exploration of low-level rules is often a challenging task, because their cardinality is commonly so large that their manual inspection becomes practically unfeasible. To overcome this issue, we early pruned redundant rules (see Section 6) and we exploited the knowledge extracted from higher-level patterns (i.e., high- and cross-level rules) to prevent experts from exploring the whole rule set. For example, given rules (1) and (8), experts may wonder what is the probability of prescribing the uric acid examination to patients who have also received a prescription for a specific liver examination. To answer this question, we can consider low-level rules (14)-(17). Specifically, their confidence values indicate the conditional probability of prescription of the uric acid examination given the occurrence of specific liver examinations in the patient dataset.

Since patient data typically contain not only examination prescriptions but also drug prescriptions, it could be also interesting to analyze the correlations between drugs (i.e., Class D-rules) and the correlations between examinations and drugs (i.e., Class ED-Rules) at different abstraction levels.

Analysis of correlations between drugs (Class D-Rules). Table 6 reports a selection of correlations between drugs. They concern the pairwise correlation between the drugs belonging to the respiratory system category (category R) and those belonging to the anti-infectives for system use category (category J). The contemporary use of drugs belonging to the above categories could prompt a detailed analysis of the corresponding guidelines [1]. Specifically, rule (3) highlights the association between the drugs belonging to the respiratory system category and drug Levofloxacin, which is commonly prescribed for infections of the respiratory system.

Analysis of correlations between examinations and drugs (Class ED-Rules). Guidelines commonly indicate established associations between examinations and drugs [1]. Their adherence could be verified against correlations between examinations and drugs mined from the real log patient data. Representative rules of this type are reported in Table 7. For example, the high-level rule (1) in Table 7 indicates a positive correlation between the examinations of the carotid and category B drugs (Blood and blood forming organs). This rule confirms the common knowledge that vascular diseases, such as problems to the carotid, are usually taken under control with drugs related to blood diseases. More specifically, the cross-level rule (2) indicates that carotid examinations are frequently associated with the category B drug with code B01AC06, which corresponds to the active principle Acetylsalicylic Acid. Acetylsalicylic Acid [7] is widely used to treat blood and vascular diseases in general (including carotid issues). Hence, the drug use appears to be coherent with guidelines. Finally, the low-level rule (3) in Table 7 shows another interesting correlation between examinations and drugs. Unlike the

former ones, it associates a specific examination (the HDL Cholesterol cardiovascular examination) with a specific drug (active principle: rosuvastatin, code: C10AA07). Rosuvastatin is indicated for cardiovascular diseases and, in particular, it is used to treat patients affected by primary hypercholesterolemia [24].

3.2.2. Profile-based correlations (Class P-Rules)

In this section we analyze the rules representing correlations between user profiles (i.e., demographic features) and treatments. These rules represent recurrences among treatments that hold for specific patient segments identified by census features (i.e., age, gender). To facilitate the analysis of different data facets, profile-based rules are further specialized into three subcategories: Age profile, Gender profile, and Age-Gender profile-based rules. A worthwhile subset of representative rules is reported in Table 8. Considering these rules allow us to characterize patients with different profiles (i.e., age and/or gender) based on their prescribed examinations and drugs.

Rules (1)-(5) in Table 8 have been classified as "Age profiles" rules (see Section 2.3), because patients are clustered into segments according to their age. For example, rule (1) indicates that diabetic patients in the age range [40-59] (i.e., middle-aged patients) are used to undergo cardiovascular examinations. The implication holds for most of the patients belonging to the segment (rule confidence 70.1%). Guidelines confirm that middle-aged diabetic patients are expected to undergo examinations in order to prevent cardiovascular diseases [1]. Furthermore, rules (2) and (3) in Table 8 indicate a positive correlation between middle-aged patients and drugs Rosuvastatin and Ramipril, respectively. Both drugs are likely to be prescribed to pa-

tients with cardiovascular diseases in conjunction with specific examinations. Drug Rosuvastatin is mainly used to treat patients with primary hypercholesterolemia, whereas Ramipril (code: C09AA05) is commonly prescribed to reduce blood pressure [7]. The confidence values of rules (2) and (3) indicate that approximately 11% of middle-aged patients actually take the specific drugs. Based on the achieved results, drug provision across medical centers and pharmacies could be shaped according to the patient age distribution. For example, medical centers that mainly treat middle-aged or elderly patients would purchase large amounts of these drugs. It is worth noticing that, to perform such analyses, discarding low-confidence rules would be harmful because they still provide information valuable for medical resource management. If we focus on middle-aged diabetic patients (age group [40-59]) to whom the HDL cholesterol examination has been prescribed at least once (see Rule (3)), then the percentage of patients who have also taken drug Rosuvastatin significantly increases with respect to all middle-aged patients (rule confidence 13.5% against 11.0%) and even the rule correlation increases (rule lift 1.82 against 1.48). Rule (6) represents a correlation between male patients and drug Finasteride. The rule appears to be reliable, because drug Finasteride is used for treatment and control of benign prostatic hyperplasia, which commonly arises in male.

3.3. Analysis of the rule distribution

The MeTA framework categorizes rules according to the covered data features and the level of abstraction of the corresponding items (see Section 2.3). We analyzed the characteristics of the rules mined using the standard configuration (i.e., *minsup*=1%, *minlift*=1.1, and *maxlength*=3).

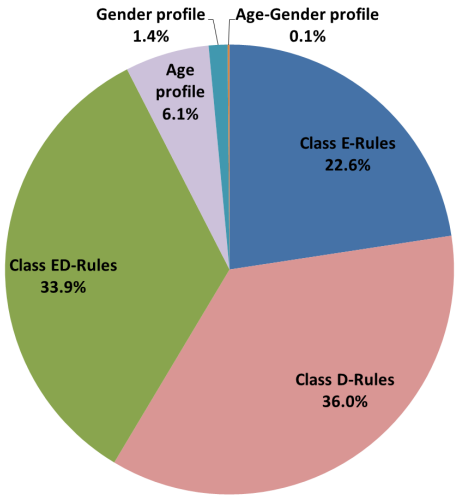

Figure 3: Percentage of rules per class. *minsup*=1%, *minlift*=1.1, and *maxlength*=3.

Figure 3 reports the rule distribution across classes E-, D-, ED-, and P-Rules. Since profile-based correlations (class P-Rules) are further classified as "Age profiles", "Gender profiles", and "Age-Gender profiles" (see Table 2 in Section 2.3) we also reported the percentage of generalized rules per subcategory. Class D-Rules and ED-Rules appear to be the largest rule sets. Both D-Rules and ED-Rules sets have a high cardinality because the number of frequently prescribed drugs in the dataset is relatively large.

We also analyzed the rule distribution across the abstraction levels of the input taxonomy. Figure 4 reports for each class the percentage of high-, cross-, and low-level rules (see Section 2.3.1). The results show that the percentage of high-level rules with respect to the total number of mined rules is always less than 2% and, for most categories, it is less than 1%. More specifically, Class E-Rules contains 55 high-level rules, Class D-Rules 238,

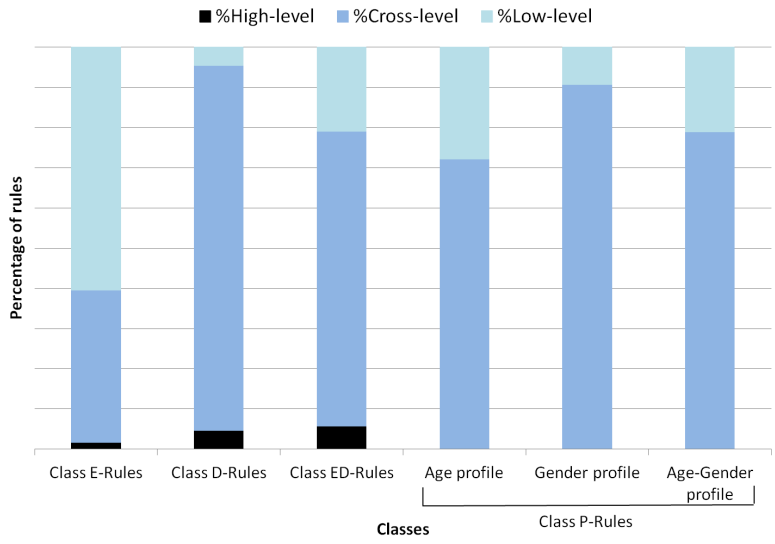

Figure 4: Percentage of rules per level. *minsup*=1%, *minlift*=1.1, and *maxlength*=3.

whereas Class ED-Rules 276. These results confirm that high-level rules provide a compact representation of the underlying correlations among data, which is particularly suitable for manual result inspection. Since the number of cross- and low-level rules is one order of magnitude higher than those of high-level rules, analyzing high-level rules first prevents experts from exploring hundreds and hundreds of (more specific) rules. For example, the experts could focus on a subset of high-level rules and then drill down to cross- and low-level rules, which represent similar information at finer granularity levels, only if high-level patterns are not informative enough to support knowledge discovery. Finally, it is worth noticing that "Age profiles", "Gender profiles", and "Age-Gender profiles" do not produce any high-level rule because no generalization hierarchy has been defined over the census features.

3.4. Quantitative analysis of the rule inspection process

MeTA addresses the issue of making the mined rule set manageable by domain experts for manual inspection by (i) selecting non-redundant rules (see Section 2.2) and (ii) performing rule exploration in a top-down fashion (see Section 2.3.1).

To analyze the pruning rate achieved by Step (i) we evaluated the ratio between the number of (traditional) frequent generalized rules and the number of (selected) non-redundant generalized rules mined from the patient dataset using the standard configuration (*minsup*=1%, *minlift*=1.1). The ratio is equal to 13.3 (i.e., on average a non-redundant generalized rule corresponds to 13.3 traditional rules) with *maxlength*=3 whereas it reaches 41.6 with *maxlength*=4, because longer rules are most likely to be redundant [48]. Since the non-redundant rule set cardinality is at least one order of magnitude smaller than the traditional one for all the tested configurations, the usefulness of the non-redundant rule selection step is confirmed.

As discussed in Section 2.3.1, rule exploration is performed in a top-down fashion starting from the most general (high-level) rules. To analyze the effectiveness of Step (ii), we analyzed both the total number of mined rules and the average number of cross- and low-level rules that experts would need to explore for each high-level rule they select during rule exploration. Table 9 reports for each category the total number of high-, cross-, and low-level rules mined using the standard configuration (i.e., *minsup*=1%, *minlift*=1.1, and *maxlength*=3). The number of high-level rules remains manageable for almost all rule categories. Among the extracted high-level rules, experts may focus on a subset according to their specific goal. For

example, based on the experts' suggestion, to plan the allocation of medical divisions according to the offered services, healthcare system managers would consider high level rules (1)-(7) in Table 5. Column (4) of Table 9 reports the ratio between the number of cross- and low-level rules and the number of high-level ones. It indicates the average number of lower-level rules per category that experts would need to explore for each high-level rule they select. The achieved results indicate that, on the average, domain experts have to explore approximately 30 cross- and low-level rules per high-level rule. Therefore, level-wise rule exploration allows the expert to perform a simplified and effective rule browsing.

We also analyzed the impact of traditional rule quality measures (i.e., support and lift) on the cardinality of the mined rule set. To this aim, we performed a large body of experiments by varying *minsup* and *minlift*, respectively, and by setting *maxlength* to 3. A separate discussion on each quality measure is given below.

Support. Decreasing the support threshold value the number of generated item combinations combinatorially increases. Hence, the support threshold significantly affects the number of extracted rules. Enforcing medium or high values may degrade the quality of the result, because some specific yet interesting rules could be discarded. Hence, we recommend users to set low support threshold values (e.g., 1%), even though this setting may generate, on average, a larger number of rules. In fact, For example, even if rule with ID 14 in Table 5 has a relatively small support value (1.4%), it was deemed to be valuable by domain experts for advanced analysis (e.g., to verify the adherence of physician's prescriptions to standard guidelines).

Lift. Enforcing lift threshold values above 1 affects the number of mined rules. As standard configuration, we enforced a minimum lift threshold equal to 1.1 on real data to prune both negatively correlated and uncorrelated item combinations. On the one hand, the interest of negatively correlated rules (i.e., rules with lift below 1) is marginal in our context of analysis. On the other hand, rules with lift close to 1 are misleading because their occurrences are not actually correlated with each other. Hence, among the positively correlated rules we further pruned approximately 10% of them because their lift value is between 1 and 1.1.

Confidence. The confidence is a commonly used rule quality measure. Unfortunately, enforcing a minimum confidence threshold could bias the quality of the mining result, because, in some cases, confidence values could be misleading [15]. Therefore, we decided against enforcing any confidence constraint.

3.5. Execution time

When coping with large or complex patient datasets, rule mining becomes the most time consuming step of the MeTA framework. Specifically, the generalized itemset mining step, driven by the support threshold, is known to be the most computationally intensive task of the rule mining process [39]. Hence, we analyzed the execution time spent by MeTA with different support thresholds. The non-redundant generalized rule mining step took less than 5 minutes while setting the minimum support threshold to 1%. When higher support thresholds are enforced, the execution time decreases super-linearly, e.g., less than 1 minutes while setting the support threshold value to *minsup*=5%.

The scalability of the proposed approach with number of transactions and transaction length is the same as traditional generalized rule mining algorithms (e.g., [39, 22, 10], i.e., it scales linearly with the dataset cardinality whereas it scales more than linearly with the number of distinct data items. In our context, the dataset cardinality corresponds to the number of considered patients, while the average transaction length strictly depends on the average number of prescribed examinations and drugs per patient.

4. Related works

Data mining techniques have largely been used to perform medical data analysis targeted to different diseases and treatments. Previous works addressed the problems of clustering (e.g., [5, 36, 4]), classifying (e.g., [25, 28]), and mining frequent patterns from healthcare data (e.g., [21]. This work addresses the analysis of a specific type of frequent patterns, i.e., the generalized association rules, mined from patient log data.

A significant research effort has been devoted to mining association rules from healthcare data. For example, sick and healthy factors for heart diseases have been investigated in [31] by exploiting three different association rule extraction algorithms, namely Apriori [3], Predictive Apriori [37], and Tertius [21]. Similarly, in [34] two of the above algorithms have been exploited to generate accurate rule-based models for type-2 diabetic patient classification. In [38], association rules have been exploited to determine two important diseases in patients diagnosed with essential hypertension, i.e., non-insulin dependent diabetes mellitus and cerebral infarction.

Parallel efforts have been devoted to taking temporal information into ac-

count during pattern mining from healthcare data. For example, the authors in [47] coupled association rule mining techniques with temporal abstraction methods to reduce hospitalization in dialysis patients, while a temporal pattern mining approach has been presented in [13] to predict the risk of developing heparin-induced thrombocytopenia. Association rules have also been applied to discover complex temporal relationships in interval-based temporal clinical data [19]. The works presented in [6, 20] addressed sequential pattern mining from healthcare data. Specifically, the authors in [6] analyzed the diagnostic pathways for colon cancer actually followed by patients, and compared them with standard medical guidelines, whereas in [20] a sequential pattern mining algorithm has been customized to manage multi-dimensional healthcare data. Unlike [13, 6, 20, 19], this work does not consider neither sequences nor temporal patterns. Conversely, it exploits generalized association rules to analyze the co-occurrences among examination/drug prescriptions at different abstraction levels regardless of the temporal order of prescriptions. To some extent, the integration of taxonomy information is complementary to both temporal and non-temporal pattern mining.

A parallel research effort has been devoted to extracting generalized association rules from potentially large datasets. Generalized association rules have first been introduced and used in [39] in the context of market basket analysis as an extension of the traditional association rule mining task [2]. The key idea was to aggregate market data items into higher-level item categories according to a user-provided taxonomy with the aim at discovering associations among data at different granularity levels. The first generalized association rule mining algorithm [39] follows the traditional Apriori-based [3]

two-step process for generalized rule mining: (i) frequent generalized itemset mining, driven by a minimum support threshold, and (ii) generalized rule generation, from the previously mined frequent itemsets, driven by a minimum confidence threshold. Candidate frequent generalized itemsets are generated by exhaustively evaluating the taxonomy. To reduce the complexity and improve the efficiency of the mining process, several algorithmic optimization strategies have been proposed (e.g., [29, 35, 39, 22]). Preliminary attempts to discover generalized patterns from medical data have been presented in [26, 14]. Specifically, in [26] the authors analyzed multiple-level co-occurrences among diseases in a public health dataset, while in [14] generalized rules are used to represent biomedical relationships between concepts occurring in Medline. With regard to the medical context, this work targets a completely different area with respect to [14, 26], i.e., it analyzes multiple-level associations among medical treatments (examinations, drugs) and patient profiles rather than among diseases or textual content of medical libraries. Concerning the performed analysis, this paper significantly improves state-of-the-art approaches, because (i) it addresses the problem of making the rule set manageable by domain experts for manual result exploration and (ii) it also considers associations among items of length above two.

When dealing with large collections of electronic health records, a huge set of patterns could potentially be generated. Hence, the readability and manageability of the mining result may significantly reduce. To overcome this issue, significant efforts have been devoted to applying pruning strategies (e.g., [12, 27, 45, 15]) on top of/in conjunction with itemset or asso-

ciation rule mining. For example, several approaches propose to push ad hoc constraints to reduce the number of mined frequent generalized itemsets [40, 22, 41, 10, 18, 16]. In the context of medical data mining association rules have been used in [8] to compactly represent correlations among examinations undergone by patients. The work in [32] proposed a new graph-based approach to reducing the cardinality of the candidate itemsets and thus discovering useful and manageable rules about medical images. In this study, we counteract the exponential growth in the number of mined generalized rules by applying an established rule pruning strategy [12] which first generates association rules on top of closed itemsets and then it prunes less informative rule extensions. Furthermore, we propose to explore rules in a top-down fashion by performing a selective drill down from most interesting high-level rules to their most specific descendant rules.

5. Conclusions and future work

This paper presents a novel approach to analyzing multiple-level correlations among medical datasets equipped with taxonomies. Since patient log dataset are often relatively sparse, discovering valuable correlations among multiple patient data features could be a challenging task. To overcome this issue, we propose to discover, categorize, and analyze non-redundant generalized association rules, which represent worthy multiple-level associations among data items.

The experiments, performed on a real diabetic patient dataset, highlight correlations among treatments and patient profiles which are consistent with the guidelines for diabetes disease [23, 1]. Furthermore, the extracted high-

level rules represent fruitful information commonly discarded by traditional rule mining approaches.

As future work, we plan to study the temporal order of examination/drug prescriptions at different abstraction levels. More specifically, we would like to extend state-of-the-art sequence and temporal pattern mining approaches (e.g., [13, 19]) by also considering taxonomy information during medical data analysis. This approach could be applicable on any patient dataset in which the temporal order of prescriptions is indicated.

An interesting development of our framework will be the application of the proposed approach to enriched patient datasets containing examination outcomes and drug posologies, because such information is strongly correlated with examination/drug prescriptions. To perform this kind of analyses a slight modification of the data representation used in the MeTA framework is needed. Furthermore, few (straightforward) preprocessing steps (e.g., data discretization) are needed prior to association rule mining. Finally, we would like to explore the applicability of the proposed approach to other contexts (e.g., genetic data [9], sports data [11], mobile data [17]) as well.

6. Acknowledgments

The authors wish to thank Dr. Baudolino Mussa and Dr. Dario Bellomo for their advices and fruitful discussions.

This work was partially supported by the GenData2020 project grant, which is funded by the Italian Ministry of Research (MIUR).

References

[1] ADA, American Diabetes Association Standards of Medical Care in Diabetes 2013, Diabetes Care 36 (2013) S11–S66.

[2] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, SIGMOD Rec. 22 (1993) 207–216.

[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp. 487–499.

[4] M.U. Ahmed, P. Funk, Mining rare cases in post-operative pain by means of outlier detection, in: 2011 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2011, pp. 35–41.

[5] D. Antonelli, E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, N.A. Mahoto, Analysis of diabetic patients through their examination history, Expert Syst. Appl. 40 (2013) 4672–4678.

[6] D. Antonelli, E. Baralis, G. Bruno, S. Chiusano, N. Mahoto, C. Petrigni, Analysis of diagnostic pathways for colon cancer, Flexible Services and Manufacturing Journal 24 (2012) 379–399.

[7] ATC, Norwegian-institute-of-public-health: Atc/ddd index 2013, 2013.

[8] E. Baralis, G. Bruno, S. Chiusano, V.C. Domenici, N.A. Mahoto, C. Petrigni, Analysis of medical pathways by means of frequent closed

sequences, in: Knowledge-Based and Intelligent Information and Engineering Systems - 14th International Conference, KES 2010, Proceedings, Part III, pp. 418–425.

[9] E. Baralis, L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, Frequent weighted itemset mining from gene expression data, in: 13th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2013, pp. 1–4.

[10] E. Baralis, L. Cagliero, T. Cerquitelli, V. D'Elia, P. Garza, Support driven opportunistic aggregation for generalized itemset extraction, in: 5th IEEE International Conference on Intelligent Systems, IS 2010, pp. 102–107.

[11] E. Baralis, T. Cerquitelli, S. Chiusano, V. D'Elia, R. Molinari, D. Susta, Early prediction of the highest workload in incremental cardiopulmonary tests, ACM TIST 4 (2013) 70.

[12] I. Batal, G.F. Cooper, M. Hauskrecht, A bayesian scoring technique for mining predictive and non-spurious rules, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012. Proceedings, Part II, pp. 260–276.

[13] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, ACM TIST 4 (2013) 63.

[14] M. Berardi, M. Lapi, P. Leo, C. Loglisci, Mining generalized association rules on biomedical literature, in: Innovations in Applied Artificial In-

telligence, 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE, pp. 500–509.

[15] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, SIGMOD Rec. 26 (1997) 265–276.

[16] L. Cagliero, Discovering temporal change patterns in the presence of taxonomies, IEEE Trans. Knowl. Data Eng. 25 (2013) 541–555.

[17] L. Cagliero, T. Cerquitelli, P. Garza, L. Grimaudo, Misleading generalized itemset discovery, Expert Syst. Appl. 41 (2014) 1400–1410.

[18] L. Cagliero, P. Garza, Itemset generalization with cardinality-based constraints, Inf. Sci. 244 (2013) 161–174.

[19] C. Combi, A. Sabaini, Extraction, analysis, and visualization of temporal association rules from interval-based clinical data, in: Artificial Intelligence in Medicine - 14th Conference on Artificial Intelligence in Medicine, AIME 2013, pp. 238–247.

[20] E. Egho, C. Raïssi, D. Ienco, N. Jay, A. Napoli, P. Poncelet, C. Quantin, M. Teisseire, Healthcare trajectory mining by combining multidimensional component and itemsets, in: New Frontiers in Mining Complex Patterns - First International Workshop, NFMCP 2012, Held in Conjunction with ECML/PKDD 2012, pp. 109–123.

[21] P. Flach, V. Maraldi, F. Riguzzi, Algorithms for efficiently and effectively using background knowledge in tertius, 2006.

[22] J. Han, Y. Fu, Mining multiple-level association rules in large databases, IEEE Trans. on Knowl. and Data Eng. 11 (1999) 798–805.

[23] ICD-9-CM, International classification of diseases, 9th revision, clinical modification, 2011.

[24] IDF, International Diabetes Federation, 2013.

[25] A.G. Karegowda, M.A. Jayaram, A.S. Manjunath, Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients, International Journal of Engineering and Advanced Technology (IJEAT) 1 (2012) 147–151.

[26] R. Kost, B. Littenberg, E.S. Chen, Exploring generalized association rule mining for disease co-occurrences, in: Proceedings of the AMIA 2012 Annual Symposium, AIMA, Chicago, Illinois, USA, 2012, pp. 1284–1293.

[27] M. Mampaey, N. Tatti, J. Vreeken, Tell me what I need to know: Succinctly summarizing data with itemsets, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 573–581.

[28] X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang, Q. Liu, Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, The Kaohsiung Journal of Medical Sciences 29 (2013) 93 – 99.

[29] J. Mennis, J.W. Liu, Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change, Transactions in GIS 9 (2005) 5–17.

[30] MeTA, MeTA source code, 2014.

[31] J. Nahar, T. Imam, K.S. Tickle, Y.P.P. Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Expert Systems with Applications 40 (2013) 1086 – 1093.

[32] H. Pan, X. Tan, Q. Han, X. Feng, G. Yin, Gma: An approach for association rules mining on medical images, in: Proceedings of the 8th International Conference on Intelligent Computing Theories and Applications, ICIC'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 425–432.

[33] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Discovering frequent closed itemsets for association rules, in: Proceedings of the 7th International Conference on Database Theory, ICDT '99, Springer-Verlag, London, UK, UK, 1999, pp. 398–416.

[34] B.M. Patil, R.C. Joshi, D. Toshniwal, Classification of type-2 diabetic patients by using apriori and predictive apriori, Int. J. Comput. Vision Robot. 2 (2011) 254–265.

[35] I. Pramudiono, M. Kitsuregawa, Fp-tax: Tree structure based generalized association rule mining, in: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD '04, ACM, New York, NY, USA, 2004, pp. 60–63.

[36] S.M. van Rooden, W.J. Heiser, J.N. Kok, D. Verbaan, J.J. van Hilten, J. Marinus, The identification of parkinson's disease subtypes using cluster analysis: A systematic review, Movement Disorders 25 (2010) 969–978.

[37] T. Scheffer, Finding association rules that trade support optimally against confidence, Intell. Data Anal. 9 (2005) 381–395.

[38] A.M. Shin, I.H. Lee, G.H. Lee, H.J. Park, H.S. Park, K.I. Yoon, J.J. Lee, Y.N. Kim, Diagnostic analysis of patients with essential hypertension using association rule mining, Healthc Inform Res 16 (2010) 77–81.

[39] R. Srikant, R. Agrawal, Mining generalized association rules, in: Proceedings of the 21th International Conference on Very Large Data Bases, VLDB '95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 407–419.

[40] R. Srikant, Q. Vu, R. Agrawal, Mining association rules with item constraints, in: D. Heckerman, H. Mannila, D. Pregibon (Eds.), Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press, Newport Beach, California, USA, 1997, pp. 67–73.

[41] K. Sriphaew, T. Theeramunkong, A new method for finding generalized frequent itemsets in generalized association rule mining, in: Proceedings of the Seventh IEEE Symposium on Computers and Communications (ISCC 2002), IEEE Computer Society, Taormina, Italy, 2002, pp. 1040–1045.

[42] P.N. Tan, V. Kumar, Interestingness measures for association patterns: A perspective, in: KDD 2000 Workshop on Post-Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics, Boston, MA, USA.

[43] P.N. Tan, V. Kumar, J. Srivastava, 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, 2002, pp. 32–41.

[44] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[45] N. Tatti, Probably the best itemsets, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 293–302.

[46] P.S. Timiras, Physiological Basis of Aging and Geriatrics, Taylor & Francis, United Kingdom, 2013.

[47] J.Y. Yeh, T.H. Wu, C.W. Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients, Decis. Support Syst. 50 (2011) 439–448.

[48] M.J. Zaki, Mining non-redundant association rules, Data Min. Knowl. Discov. 9 (2004) 223–248.

Table 5: Examples of correlations between examinations (Class E-Rules).

ID	Rule	Sup%
1	$\{(Examination, Liver)\} \rightarrow \{(Examination, Kidney)\}$	30.8%
2	$\{(Examination, Kidney)\} \rightarrow \{(Examination, Liver)\}$	30.8%
3	$\{(Examination, Liver)\} \rightarrow \{(Examination, Cardiovascular)\}$	30.8%
4	$\{(Examination, Cardiovascular)\} \rightarrow \{(Examination, Liver)\}$	30.8%
5	$\{(Examination, Kidney)\} \rightarrow \{(Examination, Cardiovascular)\}$	33.7%
6	$\{(Examination, Cardiovascular)\} \rightarrow \{(Examination, Kidney)\}$	33.7%
7	$\{(Examination, Liver), (Examination, Cardiovascular)\} \rightarrow \{(Examination, Kidney)\}$	29.3%
8	$\{(Examination, Liver)\} \rightarrow \{(Examination, Uric\ acid)\}$	24.4%
9	$\{(Examination, Liver)\} \rightarrow \{(Examination, Microscopic\ urine\ analysis)\}$	21.7%
10	$\{(Examination, Liver)\} \rightarrow \{(Examination, Culture\ urine)\}$	21.6%
11	$\{(Examination, Liver)\} \rightarrow \{(Examination, Creatinine\ clearance)\}$	16.5%
12	$\{(Examination, Liver)\} \rightarrow \{(Examination, Microalbuminuria)\}$	13.1%
13	$\{(Examination, Liver)\} \rightarrow \{(Examination, Creatinine)\}$	12.7%
14	$\{(Examination, Bilirubin)\} \rightarrow \{(Examination, Uric\ acid)\}$	1.4%
15	$\{(Examination, AST)\} \rightarrow \{(Examination, Uric\ acid)\}$	24.0%
16	$\{(Examination, ALT)\} \rightarrow \{(Examination, Uric\ acid)\}$	24.3%
17	$\{(Examination, Gamma\ GT)\} \rightarrow \{(Examination, Uric\ acid)\}$	5.3%

Table 6: Examples of correlations between drugs (Class D-Rules).

ID	Rule	Sup%	Conf%	Lift	Type
1	$\{(Drug,Category\ R)\} \rightarrow \{(Drug,Category\ J)\}$	12.5%	77.3%	1.46	High-level
	R = Respiratory system				
	J = Anti-infectives for systemic use				
2	$\{(Drug,Category\ J)\} \rightarrow \{(Drug,Category\ R)\}$	12.5%	23.7%	1.46	High-level
3	$\{(Drug,Category\ R)\} \rightarrow \{(Drug,Levofloxacin)\}$	3.5%	21.5%	1.94	Cross-level

Table 7: Examples of correlations between drugs and examinations (Class ED-Rules).

ID	Rule	Sup%	Conf%	Lift	
1	$\{(Examination,Carotid)\} \rightarrow \{(Drug,Category\ B)\}$	3%	68%	1.55	Hi
	B = Blood and blood forming organs				
2	$\{(Examination,Carotid)\} \rightarrow \{(Drug,Acetylsalicylic\ acid)\}$	2%	61%	1.94	Cr
3	$\{(Examination,HDL\ Cholesterol)\} \rightarrow \{(Drug,Rosuvastatin)\}$	3.2%	9.4%	1.26	Cr

Table 8: Examples of correlations between user profiles, drugs, and examinations (Class P-Rules).

ID	Rule	Sup%	Con
1	$\{(Age,[40\text{-}59])\} \rightarrow \{(Examination,Cardiovascular)\}$	14.8%	70.1
2	$\{(Age,[40\text{-}59])\} \rightarrow \{(Drug,Rosuvastatin)\}$	2.3%	11.0
3	$\{(Age,[40\text{-}59])\} \rightarrow \{(Drug,Ramipril)\}$	2.4%	11.6
4	$\{(Age,[40\text{-}59]), (Examination,Cardiovascular)\} \rightarrow \{(Drug,Rosuvastatin)\}$	1.82%	12.3
5	$\{(Age,[40\text{-}59]), (Examination,HDL\ Cholesterol)\} \rightarrow \{(Drug,Rosuvastatin)\}$	1.68%	13.5
6	$\{(Gender,Male)\} \rightarrow \{(Drug,Finasteride)\}$	1.0%	1.9

Table 9: Number of non-redundant rules per template.*minsup*=1%, *minlift*=1.1, *maxlength*=3.

	Number of non-redundant rules			**Average cross- and low**
Template	**#High-level**	**#Cross-level**	**#Low-level**	**rules per high-level r**
E-Rules	55	1218	1945	57.5
D-Rules	238	4652	235	20.5
ED-Rules	276	3537	1009	16.5
Age profiles	0	623	240	Not Defined
Gender profiles	0	175	18	Not Defined
Age-Gender profiles	0	15	4	Not Defined