Adverse Drug Reaction Mining in Pharmacovigilance Data Using Formal Concept Analysis

Jean Villerd[1], Yannick Toussaint[1], and Agnès Lillo-Le Louët[2]

[1] Loria – INRIA Nancy Grand Est, Nancy, France
firstname.lastname@loria.fr
[2] Pharmacovigilance Regional Center, Hôpital Européen G. Pompidou, Paris, France
agnes.lillo-lelouet@egp.aphp.fr

Abstract. In this paper we discuss the problem of extracting and evaluating associations between drugs and adverse effects in pharmacovigilance data. Approaches proposed by the medical informatics community for mining one drug - one effect pairs perform an exhaustive search strategy that precludes from mining high-order associations. Some specificities of pharmacovigilance data prevent from applying pattern mining approaches proposed by the data mining community for similar problems dealing with epidemiological studies. We argue that Formal Concept Analysis (FCA) and concept lattices constitute a suitable framework for both identifying relevant associations, and assisting experts in their evaluation task. Demographic attributes are handled so that the disproportionality of an association is computed w.r.t. the relevant population stratum to prevent confounding. We put the focus on the understandability of the results and provide evaluation facilities for experts. A real case study on a subset of the French spontaneous reporting system shows that the method identifies known adverse drug reactions and some unknown associations that has to be further investigated.

1 Introduction

Pharmacovigilance is the process of monitoring the safety of post-marketed drugs. The pharmacovigilance process starts with collecting *spontaneous case reports*: when suspecting an adverse drug reaction, health care practitioners send a case report to a spontaneous reporting system (SRS), mentioning the observed adverse effects, the drugs taken, and demographic data about the patient. These data are exploited by pharmacovigilance experts to detect *signals* of unexpected adverse drug reactions that require further clinical investigation. The size of these databases preclude their manual exploration: in 2008 more than 20,000 new cases were added to the French pharmacovigilance system while the WHO database contains more than 3 millions of reports.

The medical informatics community proposed some approaches that extract a set of *potential signals* for experts, i.e. a set of pairs (d, e) showing an unexpected correlation between an observed adverse effect e and the prescription

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 386–401, 2010.

of a marketed drug d [1,2]. Disproportionality measures have been introduced to quantify this notion of *unexpectedness* [3,4]. However, the exhaustive search strategy performed by these approaches precludes from mining high-order associations between sets of drugs and adverse effects and from efficiently applying stratification on demographic attributes to prevent confounding.

In the meantime, the data mining community introduced statistical measures from epidemiology into the itemset and rule mining problems [5,6]. Considering exposures as items and a given outcome as a class label, [7,8] proposed efficient approaches that extract *risk patterns* (or *risk itemsets*) correlated with the given outcome. The relevance of a risk pattern is measured by statistical measures such as relative risk. Efficient pruning strategies have been proposed to reduce the search space and to provide concise representations of risk patterns. In particular, [9] considered optimal risk patterns where a risk pattern is said optimal if its relative risk is greater than the relative risk of all its subpatterns. This allows to reduce the number of extracted itemsets by discarding factors that do not increase the strength of shorter risk patterns.

However, some specificities of pharmacovigilance databases compared to epidemiological studies prevent from efficiently applying the above approaches. In contrary to epidemiological studies, pharmacovigilance databases are not designed to monitor one specific exposure to a drug or one specific outcome (adverse effect). Moreover, the database only contains situations "when things went wrong", leading to many potential biases that experts should take into account. In particular, demographic features may act as confounders and lead to extract spurious potential signals. Recent studies have shown that each demographic subpopulation should be separetely investigated by performing stratification [10]. This paper deals with the following issues that are currently not addressed by available tools from the medical informatics community :

1. **Dealing with demographic factors.** Stratification is not performed on demographic factors such as age and gender because exhaustively generating measures on all strata has a prohibitive cost. The aim at dealing with demographic factors is twofold. Firstly, it provides insights into the distribution by demographic factors for a given pair (d, e) and enables a comparative study. Secondly, demographic factors are used to guide further investigation such as clinical trials, especially in patients selection.
2. **Handling complex associations.** A signal of the form (d_1, e) can be related with more complex associations involving several drugs and several adverse effects. For example, if $(d_1 d_2, e)$ is recognised as a potential drug interactions, experts should be able to compare the respective strengths of $(d_1 d_2, e)$, (d_1, e), and (d_2, e).
3. **Providing a complete information.** Since pharmacovigilance data contain many sources of bias, a potential signal (d, e) should be presented to the experts only if there is no hidden additional factor shared by the corresponding group of patients that took d and suffered from e. For instance if this subgroup only contains men, (d, e, M), i.e. (d, e) on the male subpopulation, should be rather considered. Therefore, our aim is not to find the shortest

itemsets with the highest disproportionality, but to provide experts with potential associations (D, E, X) where the itemset DEX is the most complete description of the group of patients on which the potential association is observed.

In this paper, we propose a signal detection method based on Formal Concept Analysis that provides answers to these three points. Section 2 presents the issues concerning signal detection and introduces two constraints that define potential associations. Section 3 describes our method based on a concept lattice for identifying potential associations. Section 4 presents how the concept lattice provides features that help experts in evaluating potential interactions. An experiment on real data is analysed. Section 5 concludes the paper with a summary of contributions and future work.

2 Problem Setting

Meyboom *et al.* [11] gives a comprehensive definition of signal detection process as being "A set of data constituting a hypothesis that is relevant to the rational and safe use of a medicine. Such data are usually clinical, pharmacological, pathological or epidemiological in nature. A signal consists of a hypothesis together with data and arguments." A *potential signal* is then an hypothesis suggested by an automated signal detection system that has to be evaluated by an expert. More precisely, a signal consists in (i) a pair (d, e) where d is suspected to be the cause of e (hypothesis), (ii) a set of reports (data), and (iii) disproportionality measures (arguments).

Only a few studies extended this definition to *potential associations*, i.e. higher-order hypothesis (D, E) where D and E are sets, have been published on higher-order associations, mainly about drug-drug interactions [12].

The aim of signal detection methods is to identify, among all pairs (d, e), those that occur more than expected when assuming the independance between d and e. However, although the number of reports for (d, e) is known in the database, the number of patients exposed to the drug d in the whole population is not, nor the number of patients suffering from e. Thus, the expected number of reports can not be reliably computed [13]. A solution consists in estimating the expected number of reports for (d, e) by considering the number of reports concerning other drugs and other adverse effects in the database. Therefore, contingency tables are central data structures. Table 1 depicts the contingency table for a pair (d, e). Each cell contains the number of reports corresponding to a given combination in the database: n_{11} is the number of reports containing both d and e, i.e. the observed number of reports, n_{10} is the number of reports containing d but not e, and so on. N is the total number of reports. Several measures have been introduced to capture to what extent a pair is reported more than expected. The most widely used is the *Proportional Reporting Ratio (PRR)* [3], defined as

$$PRR(d, e) = \frac{P(e|d)}{P(e|\overline{d})} = \frac{\frac{n_{11}}{n_{11}+n_{10}}}{\frac{n_{01}}{n_{01}+n_{00}}}$$

The pair (d, e) is considered to be a potential signal if $PRR \geq 2$ and $\chi^2 \geq 4$ and $n_{11} \geq 3$ [3,1]. This criterion is widely used, notably by the British Medicines and Healthcare products Regulatory Agency (MHRA). Intuitively, the first condition means that there must be twice as much probabilities to suffer from e while taking d, rather than while not taking d. The second one ensures that d and e are not independant. The third condition tells that there must be at least three reports containing d and e in the database. Other disproportionality measures such as the *Reporting Odds Ratio (ROR)* [4] are also used. More sophisticated methods implement disproportionality measures in a Bayesian framework [14].

Table 1. Contingency table for a signal (d, e)

	e	$\overline{e}$	
d	n_{11}	n_{10}	$n_{11} + n_{10}$
$\overline{d}$	n_{01}	n_{00}	$n_{01} + n_{00}$
	$n_{11} + n_{01}$	$n_{10} + n_{00}$	N

Table 2. Contingency table on a subpopulation

	eM	$\overline{e}M$	
dM	n_{11}	n_{10}	$n_{11} + n_{10}$
$\overline{d}M$	n_{01}	n_{00}	$n_{01} + n_{00}$
	$n_{11} + n_{01}$	$n_{10} + n_{00}$	$supp(M)$

Demographic factors such as gender and age may help in identifying vulnerable subpopulations. Indeed, drugs may be administered differentially according to age (e.g. vaccines), gender, or both of them (e.g. contraceptive pills), and some adverse effects may only concern a specific subpopulation (e.g. sudden infant death syndrome). Therefore, disproportionality should be computed on groups of patients that belong to the same subpopulation. This stratification process leads to compute a PRR_{strat} value on each subpopulation, called *stratum*, for a given pair (d, e). For instance, the PRR_{strat} of (d, e) on the male subpopulation is $PRR_{strat}(d, e, M) = \frac{P(e|dM)}{P(e|\overline{d}M)}$ computed from a contingency table where each cell is restricted to the male subpopulation (see Table 2 where $supp(M)$ is the number of male patients). Similarly, $\chi^2_{strat}(d, e, M)$ denotes the χ^2 value computed from the restricted contingency table. Experts compare PRR_{strat} values between strata to evaluate if the strength of the association between d and e depends on a demographic factor. For instance, if (d, e) has the same PRR_{strat} value on both male and female strata, gender is not an increasing factor.

Stratification also allows to detect situations where demographic factors act as confounders [10]. Unbalanced subpopulations may lead to situations where $PRR_{strat}(d, e, M)$ and $PRR_{strat}(d, e, F)$ are equals while $PRR_{strat}(d, e, \emptyset)$ (w.r.t. the whole population) has a different value. In such case, $PRR_{strat}(d, e, \emptyset)$ is not reliable and is said to be counfounded by gender. Therefore both *crude* $PRR_{strat}(d, e, \emptyset)$ and $PRR_{strat}(d, e, x_i)$ on strata x_i are relevant for experts to evaluate the strength and the reliability of a signal (d, e).

The three initial issues mentioned in introduction can be refined in extracting potential associations (D, E, X) such that:

1. the disproportionality of (D, E, X) is computed w.r.t. the subpopulation X, following the stratification strategy;

2. potential associations are presented to the experts in such a way that comparisons between an association (d_1, e, M) and its related associations (e.g. $(d_1 d_2, e, M)$, $(d_1, e, \emptyset)$) is straightforward;
3. considering a potential association (D, E, X), the corresponding group of patients do not share any additional attribute than those in DEX.

3 A FCA-Based Signal Detection Method

Let $\mathcal{D}$ be a set of drugs, $\mathcal{E}$ be a set of adverse effects and $\mathcal{X}$ a set of binarized demographic attributes. We look for potential associations (D, E, X), ($D \subseteq \mathcal{D}$, $E \subseteq \mathcal{E}$, $X \subseteq \mathcal{X}$, $D \neq \emptyset$, $E \neq \emptyset$) that satisfy two types of constraints:

- a closure constraint: stating that patients that cover the itemset $D \cup E \cup X$, noted DEX, do not share any additional attribute,
- a strength constraint: stating that $supp(DEX) \geq 3$, $PRR_{strat}(D, E, X) \geq 2$ and $\chi^2_{strat}(D, E, X) \geq 4$.

The closure constraint clearly says that DEX must be a closed itemset. Thus, our search space for potential associations consists of closed itemsets that contain at least one element of $\mathcal{D}$ and one element of $\mathcal{E}$. In the following we present basics on Formal Concept Analysis and concept lattices. We later show that the concept lattice is a suitable structure for extracting potential associations, in the sense that it covers our search space, and that it provides experts with efficient ways of comparing related associations.

3.1 Basics on Formal Concept Analysis

Considering a binary relation between a set of objects $\mathcal{O}$ and a set of binary attributes $\mathcal{A}$, FCA extracts a set of pairs (O, A) with $O \subseteq \mathcal{O}$, $A \subseteq \mathcal{A}$, called formal concepts, such that each object in O owns all attributes in A and vice-versa. Formal concepts are partially ordered w.r.t. the inclusion of O and A, to form a lattice structure called concept lattice. In that way, the concept lattice can be seen as a conceptualization of the binary relation.

In the following, we present formal definitions from [15]. A *formal context* is a triple $\mathbb{K} = (\mathcal{O}, \mathcal{A}, I)$ where $\mathcal{O}$ is a set of *objects*, $\mathcal{A}$ a set of *attributes*, and $I \subseteq \mathcal{O} \times \mathcal{A}$ a binary relation such that oIa if the object o owns the attribute a. Figure 1 shows a formal context $\mathbb{K}$ with $\mathcal{O} = \{o_1 \dots o_7\}$ and $\mathcal{A} = \{d_1 \dots d_3\} \cup \{e_1, e_2\} \cup \{M, F\}$.

Two *derivation operators*, both denoted by $(.)'$, link objects and attributes. Considering a set of objects $O \subseteq \mathcal{O}$, $O' = \{a \in \mathcal{A} | oIa\}$, i.e. O' is the set of attributes shared by all objects in O. Dually, $A' = \{o \in \mathcal{O} | oIa\}$ is the set of objects that own all attributes in A. $|A'|$ is called the *support* of A, noted $\sigma(A)$. For instance, $\{d_1, d_2\}' = \{o_3, o_4\}$ and $\{o_3, o_4\}' = \{d_1, d_2, e_1, M\}$.

Two compound operators, both denoted by $(.)''$, composed of the two previous derivation operators, are *closure operators* on $2^{\mathcal{O}}$ and $2^{\mathcal{A}}$. Therefore O'' is the maximal set of objects that share the same attributes than the objects in O.

Dually, A'' is the maximal set of attributes that are owned by the objects that share attributes in A. A set of attribute A is said to be *closed* if $A = A''$. The set of sets B such that $B'' = A$ forms the *equivalence class* of A. All sets in the equivalence class of A have the same support $\sigma(A)$. For instance, $\{d_1, d_2\}$ is not closed since $\{d_1, d_2\}'' = \{o_3, o_4\}' = \{d_1, d_2, e_1, M\}$, while $\{o_3, o_4\}$ is closed since $\{o_3, o_4\}'' = \{d_1, d_2, e_1, M\}' = \{o_3, o_4\}$.

A *formal concept* is a pair (O, A) such that $O = O''$ and $A = O'$. Each object in O owns all attributes in A and vice-versa. Both O and A are closed sets, which means that no object (resp. attribute) can be added to O (resp. A) without changing A (resp. O). O (resp. A) is called the *extent* noted $\text{Ext}(O, A)$ (resp. the *intent* noted $\text{Int}(O, A)$) of the concept. The set of all formal concepts of the formal context $\mathbb{K}$ is denoted $\mathfrak{B}(\mathbb{K})$. For instance, $(\{o_3, o_4\}, \{d_1, d_2, e_1, M\})$ is a formal concept.

Formal concepts are partially ordered w.r.t. to the inclusion of their extents. Considering two concepts (O_1, A_1) and (O_2, A_2), $(O_1, A_1) \leq (O_2, A_2)$ iff $O_1 \subseteq O_2$ (which is equivalent to $A_1 \supseteq A_2$). The set of all formal concepts ordered in this way is denoted by $\underline{\mathfrak{B}}(\mathbb{K})$ and is called the *concept lattice* of the formal context $\mathbb{K}$. The maximal concept $(\mathcal{O}, \mathcal{O}')$ is called the *top* concept, and the minimal concept $(\mathcal{A}', \mathcal{A})$ is called the *bottom* concept.

The concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$, built from $\mathbb{K}$ is shown in Figure 1. Each box represents a formal concept with its intent in the upper part, and its extent in its lower part.

Considering an attribute a, its *attribute concept*, denoted $\mu(a)$, is the unique concept (a'', a'), i.e. the highest concept that contains a in its intent on Figure 1. For instance, $\mu(e_2) = (\{o_1, o_7\}, \{e_2\})$.

In the worst case, the number of concepts of $\mathbb{K} = (\mathcal{O}, \mathcal{A}, I)$ is $2^{\min(|\mathcal{O}|, |\mathcal{A}|)}$. This occurs when each subset of $\mathcal{O}$ or $\mathcal{A}$ is closed, which is improbable in practice.

3.2 Our Approach

Our aim is to extract potential associations that satisfy a closure constraint and a strength constraint. We showed that only closed itemsets can satisfy these constraints. Moreover our aim is to provide an understandable representation of results. As said before, interpretation is a difficult task for experts since pharmacovigilance data may contain many biases. Since the content of the database is not the result of a sampling method, spurious potential associations may be extracted. Disproportionality measures can not make the difference between a spurious disproportion due to a selection bias and a real disproportion due to an adverse effect reaction. Only experts can make this difference w.r.t. the content of the database and their domain knowledge. Therefore, in order to evaluate a potential association (D, E, X), experts need more information than disproportionality measures. They need to put back the association in its context of extraction, i.e. in the portion of the database where the disproportionality occurs.

The concept lattice is then a suitable structure for signal detection. It is built from the context $(\mathcal{O}, \mathcal{A}, I)$, where $\mathcal{O}$ is the set of reports, and $\mathcal{A} = \mathcal{D} \cup \mathcal{E} \cup \mathcal{X}$ is the set of attributes. Since concept intents are closed itemsets, the search

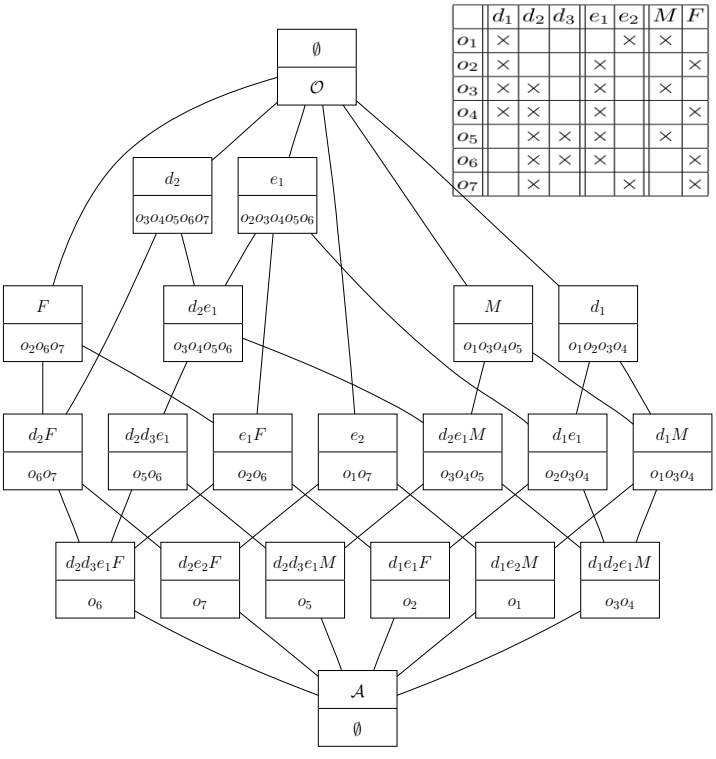

Fig. 1. A formal context and its associated concept lattice

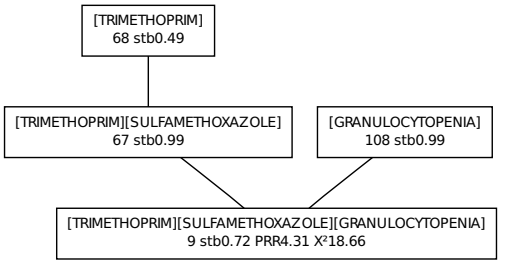

Fig. 2. An interaction example containing noise

space for potential associations is the set of concepts. Moreover, considering an association (d_1d_2, e_1, M), the partial order between concepts allows to isolate relevant information for interpretation that will be presented to experts: more specific strata among descendants of the concept with intent $\{d_1, d_2, e, M\}$, more general strata among ascendants for instance. But also to compare strengths of related associations: more specific associations (e.g. (d_1d_2, e_1e_2, M)) will be found among descendants and more general among ascendants.

Thus, our algorithm for extracting potential associations is straightforward. Concepts whose intent contains at least one drug and one adverse effect, and whose extent contains at least three reports are considered as candidate associations. Their contingency table is computed w.r.t. the demographic attributes in intent. If the MHRA criterion is satisfied, the intent is added to the set of potential associations.

Data: a concept lattice $\mathcal{L}$
Result: a set of potential associations P
foreach *concept* $c \in \mathcal{L}$ **do**
  if *Int(c) contains at least one element of* $\mathcal{D}$ *and one element of* $\mathcal{E}$ *and* $|Ext(c)| \geq 3$ **then**
    compute the contingency table for $\text{Int}(c)$
    compute PRR_{strat} and χ^2_{strat} values from the contingency table
    if $PRR_{strat} \geq 2$ *and* $\chi^2_{strat} \geq 4$ **then**
      add $\text{Int}(c)$ to the set of potential associations P
    end
  end
end

Algorithm for signal detection

Therefore, the number of candidate associations is bounded by $2^{\min|\mathcal{O}|,|\mathcal{A}|}$, which is the number of formal concepts in the worst case. In practice, the number of reports is larger than the number of attributes, and all subsets of $\mathcal{A}$ are not closed.

Computing contingency tables. Contingency tables are built from the lattice, in order to compute PRR_{strat} and χ^2_{strat} values.

Since each candidate association (D, E, X) is a closed itemset, there exists a unique formal concept c_{DEX} with $\text{Int}(c_{DEX}) = D \cup E \cup X$. We show in the following that the contingency table of any association can be computed knowing the support of c_{DEX} and the extent of the attribute-concepts $\mu(a), a \in DEX$.

In the general case of an association (D, E, X), The cell values of its contingency table restricted to the subpopulation X are computed as follows.

$$n_{11} = \sigma(DEX) = |\bigcap_{a \in DEX} \text{Ext}(\mu(a))| = |\text{Ext}(c_{DEX})|$$

$$n_{10} = \sigma(D\overline{E}X) = \sigma(DX) - \sigma(DEX) = |\bigcap_{a \in DX} \text{Ext}(\mu(a))| - n_{11}$$

$$n_{01} = \sigma(\overline{D}EX) = \sigma(EX) - \sigma(DEX) = |\bigcap_{a \in EX} \text{Ext}(\mu(a))| - n_{11}$$

$$n_{00} = \sigma(\overline{D}\,\overline{E}X) = |\bigcap_{a \in X} \text{Ext}(\mu(a))| - (n_{11} + n_{10} + n_{01})$$

Insights for noise detection. The concept lattice provides an additional measure that helps in evaluating the reliability of a potential association. The *stability index* of a formal concept [16] quantifies the ability of the concept to remain existent after deletion of objects in its extent. In other words, the stability index of a concept c is low if $\text{Int}(c)$ becomes non-closed after the removal of a few objects from $\text{Ext}(c)$. Then, an unstable concept c correspond to a *barely closed* itemset $\text{Int}(c)$. Therefore, stability can be presented to experts as an additional quality measure for potential association. A potential association (D, E, X) with a low stability index barely satisfies the closure constraint and should be considered with care by experts.

Moreover, stability can provide insights for detecting noisy reports. We illustrate this aspect on a real example. Trimethoprim (d_1) and sulfamethoxazole (d_2) come together in the dosage form of marketed drugs, thus a unique concept $\mu(d_1) = \mu(d_2)$ should exist in the lattice. It is not the case (see Figure 2) since $\mu(d_2) \leq \mu(d_1)$ and $\sigma(\mu(d_1)) = 68$ while $\sigma(\mu(d_2)) = 67$. This means that, among all patients that took d_1, only one did not take d_2, which probably corresponds to a badly filled report. The stability index can capture such a situation. Here, $\mu(d_1)$ has a low stability since the removal of the noisy report will lead d_1 to become non-closed with $d_1'' = \{d_1, d_2\}$ and then $\mu(d_1)$ will become $\mu(d_1) = \mu(d_2)$. Thus, a low stability index for a given concept should draw experts' attention to the potentially noisy reports contained in its extent.

3.3 Related Work

Several works focused on finding *risk patterns* in epidemiological studies. Considering a set of patients described by a set of nominal attributes, and a target outcome e that partitions patients into two classes (presence/absence), a risk pattern is a set of attribute-value pairs D such that the pattern is locally frequent ($support(De) \geq min\_sup$) and its *relative risk* is higher than a given threshold. Relative risk $RR(D, e) = \frac{P(e|D)}{P(e|\overline{D})}$ is a widely used measure in epidemiological studies. Note that PRR and RR formula are identical when ignoring demographic factors. [9] proposed algorithms for efficiently mining risk patterns. A risk pattern is said *optimal* if its relative risk is greater than the relative risk of all its subpatterns. This allows to reduce the number of extracted patterns by discarding factors that do not increase the strength of more general risk patterns.

Although PRR and RR formula are identical for a given outcome e and a set of attributes D, this approach does not fit well our requirement for pharmacovigilance.

First there is no predefined outcome in pharmacovigilance data. Each combination of adverse effects may be considered as an outcome. Applying the precited approach would consist in generating the set of optimal risk patterns for each combination of adverse effects. This also prevents from applying other approaches such as subgroup discovery [17] and contrast set mining [18].

Secondly, in [9] demographic attributes play the same role as drugs in contigency tables. This means that the PRR of the pattern $\{d, M\}$ is computed as

$PRR(d, e, M) = \frac{P(e|d,M)}{P(e|\overline{d},M)}$. In order to be consistent with the stratification recommendation about demographic factors, the PRR of $\{d, M\}$ should be computed w.r.t. the male stratum. It should compare men that took d and suffered from e with men that did not take d and suffered from e: $PRR_{strat}(d, e, M) = \frac{P(e|d,M)}{P(e|\overline{d},M)}$.

Thirdly, as defined in [9], risk patterns may not be closed itemsets and therefore may not satisfy our closure constraint.

Suppose that d_1d_2e is a closed itemset and that the closure of d_1 is $d_1'' = d_1d_2$, then $PRR(d_1, e, \emptyset) = PRR(d_1d_2, e, \emptyset)$ as well as $PRR_{strat}(d_1, e, \emptyset) = PRR_{strat}(d_1d_2, e, \emptyset)$. The risk pattern d_1d_2 is not optimal since its PRR value is not higher than its subpattern d_1 and is not retrieved, while following our constraints d_1d_2 has to be retrieved and not d_1. Moreover, the fact that risk patterns are extracted w.r.t. a given outcome would lead to generate risk patterns for e_1 and then risk patterns for e_2 without paying attention to situations where $e_1'' = e_2$. In this case, risk patterns w.r.t. e_1 do not satisfy our closure constraint since all patients that suffer from e_1 also suffer from e_2.

Moreover, considering non-closed itemsets prevents from computing PRR_{strat} in an accurate way. Consider the group of patients that took a drug d. Suppose that all patients that took d are men, i.e. the closure of $\{d\}$ is $\{d, M\}$. Then $PRR_{strat}(d, e, \emptyset) = \frac{P(e|d,\emptyset)}{P(e|\overline{d},\emptyset)} = \frac{P(e|d,M)}{P(e|\overline{d},\emptyset)}$. The numerator group of patients actually belongs to a more specific stratum (men) than the denominator group (men and women). $PRR_{strat}(d, e, \emptyset)$ can not be reliably computed w.r.t. the available data since only men took d. In this case, associations involving d are only reliable w.r.t. the male subpopulation. No reliable hypothesis can be made about d and e on the whole population since there is no female subpopulation that would allow to evaluate if (d, e) depends on gender or not.

Since signal detection aims at providing experts with hypothesis for further investigation, we claim that the reliability of an hypothesis is at least as important as its statistical strength. An hypothesis (D, E, X) is reliable if the corresponding set of patients do not share an additional attribute that may delude experts, i.e. if DEX is a closed itemset. This is true for demographic attribute as shown before but also for drugs and adverse effects.

4 Evaluation Facilities and Experimentation

This section shows how experts get a *contextualized* association using our approach. In addition to disproportionality measures, insights are given to help them in deciding whether a signal or an interaction should be further investigated or not.

4.1 Visualization and Navigation

The core idea is to use the concept lattice as a synthetic representation of the database. From the list of potential association, experts access to a detailed

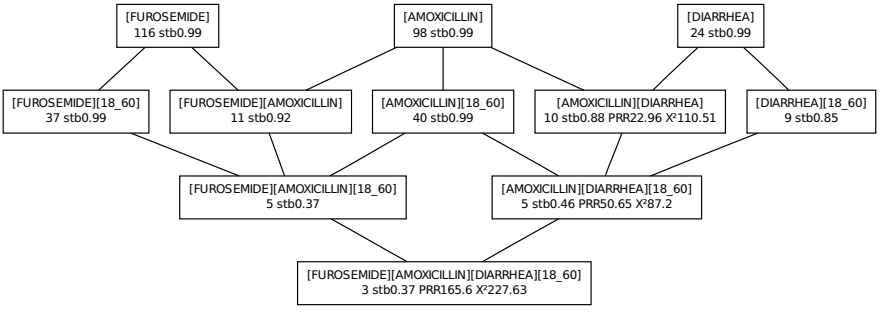

Fig. 3. Subpart of the lattice illustrating a potential interaction

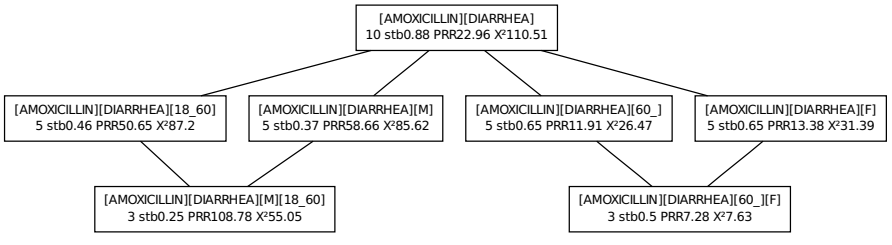

Fig. 4. Comparison of the different strata of a potential signal

view that shows a subpart of the concept lattice, revealing additional information compared to statistical measures and helping experts in their interpretation and evaluation task.

Figure 3 shows the user interface illustrating a potential interaction $(d_1 d_2, e, X)$ where d_1 is `amoxicillin`, d_2 is `furosemide`, e is `diarrhea` and X is `18_60`, meaning age between 18 and 60.

A subpart of the lattice is shown, which contains the concept $c_{d_1 d_2 e X}$, corresponding to the interaction, at the bottom, the attribute-concepts $\mu(d_1)$, $\mu(d_2)$, $\mu(e)$ at the top, and all concepts on the paths from $c_{d_1 d_2 e X}$ to the attribute-concepts. Then the graph shows concepts that are more general than $c_{d_1 d_2 e X}$.

Concepts are labeled with their intent, support and stability. Concepts that own at least one drug and one adverse effect are also labelled with PRR_{strat} and χ^2_{strat} values. Through this graph, experts can compare the PRR_{strat} values of the interaction $(d_1 d_2, e, X)$ with those of the signals (d_1, e, X) and $(d_1, e, \emptyset)$, and observe that there are no concepts representing the signals (d_2, e, X) and $(d_2, e, \emptyset)$. This gives the information that no patient took `furosemide` and suffered from `diarrhea` without `amoxicillin`. Concepts that do not correspond to associations are also relevant. For instance, experts can observe that among the 24 patients that suffered from `diarrhea`, 10 took `amoxicillin` and state whether this ratio is realistic or is due to a selection bias.

Another graph (cf. Figure 4) shows a given association as root and those of its subconcepts that correspond to its demographic strata with a least 3 reports. Experts can compare their respective PRR_{strat} values and observe that, in this example, age distribution is different in male and female strata.

4.2 Experimentation

We applied our method on a subset of the French national SRS database. This subset contains 3249 cases, 976 drugs, 573 adverse effects. Two demographic attributes, gender and age are binarized into 6 binary attributes (2 for gender and 4 for age). The resulting lattice contains 13178 concepts, among which 6788 with support ≥ 3. Since only signals (one drug, one adverse effects), and interactions (two drugs, one adverse effect) are currently considered by pharmacovigilance experts, we only showed potential signals and interactions to experts. The 2812 candidate signals led to 786 potential signals and the 836 candidate interactions to 183 potential interactions.

Review of potential signals. Potential signals were reviewed by an expert who classified them into 5 categories (see Table 3). Categories (1),(2) contain true positives, (3),(4) false positives and (5) unknown potential signals. 27 signals were classified as unknown, i.e. not reported in the literature, but interesting enough for further investigation by experts.

True positives are consistent with results of previous studies [19] and no known true-positive is missing. In the majority of cases, the demographic attributes associated to the couple drug/effect constitute a known risk factor or probable risk factor. For example, cases of `Pulmonary Hypertension` associated with the use of appetite suppressants amphetamine-like were observed in women, between the ages of 18 and 60.

False positives (contained in categories (3) and (4)) are common in signal detection and some of them are well-known. The signal (`hydrochlorothiazide`, `cough`) is detected because these drug and adverse effects often appear together. However in these cases, cough is actually caused by ACE inhibitors taken concomitantly with `hydrochlorothiazide`. Since there are several ACE inhibitors d_i, each association (d_i,`cough`) appears with a lower support than the association (`hydrochlorothiazide`, `cough`), which may delude experts. A solution would be to introduce drug therapeutic families, such as ACE, as attributes, with (o, `ACE`) $\in I$ for each case o containing an ACE inhibitor. Then signals of the form (`ACE`, `cough`) would be detected, where ACE is a drug family, even if each signal (d, e) where d is an ACE inhibitor is too rare to be detected. Current improvements of our method aim at solving this problem.

Review of potential interactions. The evaluation of interactions is more difficult since it involves complex pharmacokinetics aspects. Moreover there is no consensus on whether (d_1d_2, e, X) should be considered as an interaction when both d_1 and d_2 are known to be the cause of e. Thus, we are not able to separate

Table 3. Potential signals

category	count	
1. known (in reference documents)	720 (91.6%)	true positives
2. known (in a similar form)	24 (3.1%)	
3. the effect is the origin of the medication	3 (0.4%)	false positives
4. due to concomitant drug	11 (1.4%)	
5. unknown potential signal	28 (3.5%)	further investigations needed

Table 4. Potential interactions

category	count
either d_1 or d_2 is a known cause of e	64(35.0%)
both d_1 or d_2 are known causes of e	66(36.0%)
d_1 and d_2 in the same dosage form	34(18.6%)
neither d_1 or d_2 are known causes	19(10.4%)

true and false positives. Experts classified the 183 potential interactions into 4 categories (see Table 4). The last category corresponds to cases where further investigations are needed.

We noted that, in some cases, the PRR_{strat} value of an interaction (d_1d_2, e, X) where only d_1 is a known cause of e was greater than $PRR_{strat}(d_1, e, X)$. In such cases, it is not clear if the focus should be put on (d_1d_2, e, X) or on (d_1, e, X). To our knowledge, there has been no pharmacovigilance study on defining preferences between an interaction (d_1d_2, e, X) and a signal (d_1, e, X) w.r.t. PRR value. Therefore, we can not discard (d_1d_2, e, X) when $PRR_{strat}(d_1d_2, e, X) < PRR_{strat}(d_1, e, X)$. This prevents from using the pruning strategy of the optimal risk patterns approach [9], that would discard (d_1d_2, e, X).

Detection of noisy reports. In a previous section, we showed that the stability index of a concept may be a clue for noisy reports detection. However, we faced the difficulty of defining a threshold on stability that defines *unstable* concepts. Frequent unstable concepts are interesting. They can be seen as concepts that gather a high number reports, but that actually exist because of only a few of them, which may be noisy reports. Frequent unstable concepts should be found in the upper left hand corner of the Figure 5. We empirically decided to investigate the 20 concepts with a minimum support of 20 reports and a stability index below 0.5. All of these concepts were in the same configuration than in Figure 2, i.e. among the n reports gathered by the unstable concept, $n-1$ also share another attribute. For instance, among the 20 reports gathered by the unstable concept with intent {`tacrine`, `M`}, 19 also own the attribute $age > 60$. Since tacrine is used in the treatment of Alzheimer's disease, the report that does not own $age > 60$ is suspect and should be verified. The expert considered that the n^{th} report was actually suspect in 19 of the 20 unstable concepts under review.

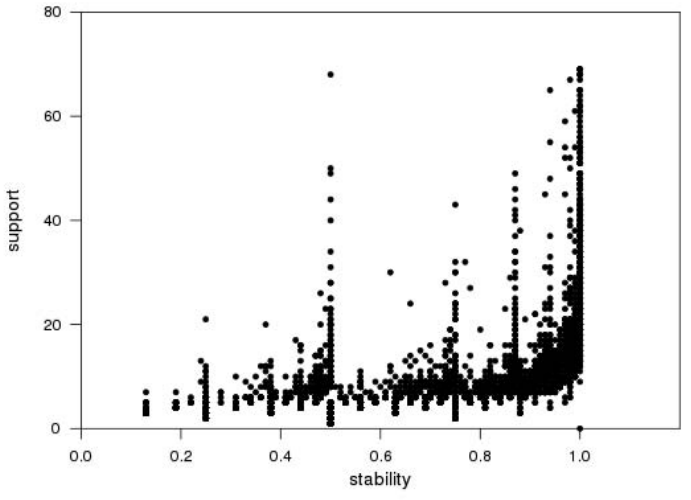

Fig. 5. Stability and support

5 Conclusion

In this paper, we presented an automated signal detection method, based on concept lattices, that provides a framework for extracting potential associations and performing qualitative analysis of the extracted associations.

We claim that only associations that are closed itemsets should be presented to experts, since non-closed associations do not fully describe the set of factors shared by a subgroup of patients. Demographic attributes are taken into account in the PRR computation so that the disproportionality of an association is computed w.r.t. the subpopulation in which the association is observed. The closure constraint allows to identify the accurate subpopulations and prevents from exhaustively evaluating each population stratum.

Our method is thought for extracting complex associations, i.e. extracting associations where there are one or more drugs, one or more adverse effects and several demographic factors. Nowadays, if signals have been quite well studied, little work has been done on interactions, and practically none on syndromes (1 drug, several effects) or protocoles (several drugs, several effects) which justifies the facts that our evaluation has only been performed on signal and interactions.

When evaluating extracted associations, experts have access to subparts of the lattice for visualizing related associations, for example, an interaction is displayed with its related signals as well as its different "strengths" on subpopulations. This visualization is of particular interest when both a signal and an interaction pass the MHRA criterion. Only experts – no automated process – are able to decide which of signals and interactions should be validated, mostly because of pharmacokinetics complexity. The interface is designed to facilitate a qualitative analysis by experts and guides exploration, interpretation and validation of associations.

Our approach is closely related to *domain driven data mining* [20]: starting from a domain-specific problem, our goal is to discover actionable knowledge to satisfy user needs. Here actionable knowledge consists of unexpected associations that need further investigations. These *actionable* associations are identified by experts among *interesting* associations that satisfy strength and closure constraints. The interface supports experts in finding actionable associations among interesting ones. This approach could be used for other applications where a synthetic graphical view is needed by experts to evaluate the actionability of extracted patterns. Finally, we are currently investigating solutions that include domain knowledge such as families of drugs and adverse effects.

References

1. Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L., Puijenbroek, E.P.V.: The role of data mining in pharmacovigilance. Expert Opinion on Drug Safety 4(5), 929–948 (2005)
2. Bate, A., Lindquist, M., Edwards, I.R.: The application of knwoledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundamental & Clinical Pharmacology 22(2), 127–140 (2008)
3. Evans, S.J.W., Waller, P.C., Davis, S.: Use of proportional reporting ratios for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and Drug Safety 10(6), 483–486 (2001)
4. van der Heijden, P.G.M., van Puijenbroek, E.P., van Buuren, S., van der Hofstede, J.W.: On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. Statistics in Medicine 21(14), 2027–2044 (2002)
5. Morishita, S., Sese, J.: Transversing itemset lattices with statistical metric pruning. In: Proc. of the 19th ACM Symposium on Principles of Database Systems, pp. 226–236. ACM, New York (2000)
6. Gu, L., Li, J., He, H., Williams, G., Hawkins, S., Kelman, C.: Association rule discovery with unbalanced class distributions. In: Gedeon, T(T.) D., Fung, L.C.C. (eds.) AI 2003. LNCS (LNAI), vol. 2903, pp. 221–232. Springer, Heidelberg (2003)
7. Li, H., Li, J., Wong, L., Feng, M., Tan, Y.: Relative risk and odds ratio: A data mining perspective. In: Proc. of the 24th ACM Symposium on Principles of Database Systems, p. 377. ACM, New York (2005)
8. Li, J., Fu, A., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R., Kelman, C.: Mining risk patterns in medical data. In: 11th ACM International Conference on Knowledge Discovery in Data Mining, pp. 770–775. ACM Press, New York (2005)
9. Li, J., Fu, A., Fahey, P.: Efficient discovery of risk patterns in medical data. Arificial Intelligence in Medicine 45(1), 77–89 (2009)
10. Woo, E., Ball, R., Burwen, D., Braun, M.: Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System. Drug safety 31(8), 667–674 (2008)
11. Meyboom, R.H., Egberts, A.C., Edwards, I.R., Hekster, Y.A., De Koning, F.H.P., Gribnau, F.W.J.: Principles of signal detection in pharmacovigilance. Drug Safety 16(6), 335–365 (1997)

12. Almenoff, J., DuMouchel, W., Kindman, L., Yang, X., Fram, D.: Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiology and Drug Safety 12(6), 517–521 (2003)
13. Roux, E., Thiessard, F., Fourrier, A., Bégaud, B., Tubert-Bitter, P.: Evaluation of statistical association measures for the automatic signal detection generation in pharmacovigilance. IEEE Transactions on Information Technology in Biomedicine 9(4), 518–527 (2005)
14. DuMouchel, W.: Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. The American Statistician 53(3), 177–190 (1999)
15. Ganter, B., Wille, R.: Formal concept analysis: Mathematical Foundations. Springer, Berlin (1999)
16. Kuznetsov, S.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49(1-4), 101–115 (2007)
17. Boley, M., Grosskreutz, H.: Non-redundant Subgroup Discovery Using a Closure System. In: Proc. of ECML/PKDD, p. 194. Springer, Heidelberg (2009)
18. Bay, S., Pazzani, M.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3), 213–246 (2001)
19. Bousquet, C., Sadakhom, C., Le Beller, C., Jaulen, M., Lillo-Le Louët, A.: Revue des signaux générés par une méthode automatisée sur 3324 cas de pharmacovigilance. Thérapie 61(1), 39–47 (2006)
20. Cao, L., Zhang, C., Yu, P.S., Zhao, Y.: Domain Driven Data Mining. Springer, New York (2010)