MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-969

A Qualitative Approach to Signal Mining in Pharmacovigilance using Formal Concept Analysis

Agnès Lillo-Le Louët[a], Yannick Toussaint[b], Jean Villerd[b]

[a] *Centre régional de pharmacovigilance, Hôpital européen Georges Pompidou, Paris, France*
[b] *LORIA – INRIA Nancy Grand Est, Vandœuvre-lès-Nancy, France*

Abstract

"Pharmacovigilance is the process and science of monitoring the safety of medicines, consisting in (i) collecting and managing data on the safety of medicines (ii) looking at the data to detect 'signals' (any new or changing safety issue)" [1]. Pharmacovigilance is mainly based on spontaneous reports: when suspecting an adverse drug reaction, health care practitioners send a report to a spontaneous reporting system (SRS). This produces huge databases containing numerous reports and their manual exploration is both cost and time prohibitive. Existing techniques that automatically extract relevant signals rely on statistics or Bayesian models but do not provide information to the experts about possible biases lying in the data, nor about the specificity of a signal to a particular patient profile. Our extraction method combines numerical methods from the state of the art with a qualitative approach that helps interpretation. We build a synthetic representation of the database that is used to (i) identify unexpected patterns and biases (ii) extract potentially relevant signals w.r.t. patient profiles (iii) provide traceability facilities between extracted signals and raw data.

Keywords:

Adverse drug reaction reporting systems, Data interpretation, Data collection, Drug safety, Public health informatics, Information storage and retrieval, Artificial intelligence, Formal concept analysis, Data mining

Introduction

The huge and constant increasing size of spontaneous reporting systems (SRS) precludes case-by-case human analysis. Indeed, in 2008 more than 20,000 new cases were added to the French pharmacovigilance system; the WHO database contains more than 3 millions of reports. This has lead to the development of data mining algorithms (DMAs) that automatically extract signals, i.e. potentially relevant adverse drug reactions for further investigations by experts in pharmacovigilance [2].

Two main approaches to extract signals are known, both of them are based on statistical criteria: (i) the frequentist approach which establishes pertinence threshold with respect to disproportionality measures between occurrences of a drug and an adverse effect (AE), and (ii) the Bayesian approach based on probability distribution models. Most of the debate has focused on the advantages and drawbacks of these approaches and on the fine tuning of their respective measures and thresholds.

DMAs only deal with co-occurrence of drugs and AE and produce quantitative indicators. For this reason, [3] throws them back into question arguing that a drug-AE pair is, in itself, rarely sufficient to assess whether a potential signal has been generated. Indeed, using DMAs, experts have to evaluate each extracted drug-AE pair and its statistical measures with no way to estimate to what extent these measures are reliable. They also ignore if there are some demographic population restrictions or the presence of concomitant medications.

Moreover some biases in SRS databases degrade quality of the DMA results: the number of patients that take a particular drug without AE is not known (no control sample) and fields in the database are not always fully or properly filled. Experts ignore if some of the detected signals are due to biases since they have no way to evaluate the presence of specific biases or noise in the case database when they evaluate detected signals.

We argue that DMAs should provide experts disproportionality measures as well as qualitative information in order to explain or to trace the reasons why each signal has been generated. We claim that a symbolic classification method such as Formal Concept Analysis reaches this goal providing (i) a synthetic view of the database (ii) a search space for candidate signals (iii) an environment to navigate among results (iv) and a potential noise detection method.

Materials and Methods

This section presents first Formal Concept Analysis which builds a partial ordered structure called lattice. Then, we highlight some mathematical properties of the lattices used to achieve the (i) to (iv) previous goals.

Table 1 – Binary relation between objects in rows (cases) and attributes in columns (age bracket, gender, AE, drugs)

	demographic attributes					**adverse effects**										**drugs**				
	<18	**18-60**	**>60**	**M**	**F**	**e1**	**e2**	**e3**	**e4**	**e5**	**e6**	**e7**	**e8**	**e9**	**e10**	**d1**	**d2**	**d3**	**d4**	**d5**
#1		×			×				×		×				×		×			×
#2			×	×			×		×	×	×	×	×				×	×		
#3		×			×	×	×		×				×	×					×	×
#4	×				×	×	×						×				×	×		
#5		×		×		×		×			×	×		×	×	×				

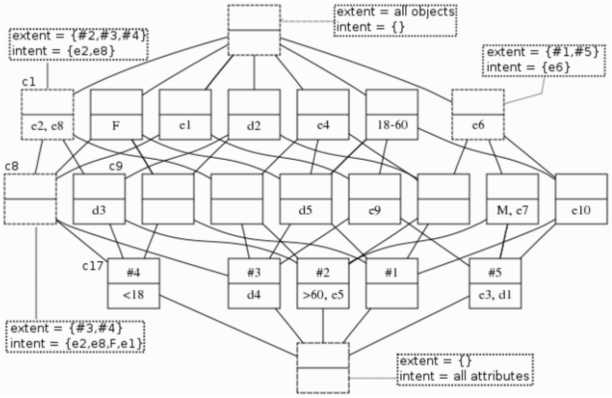


Figure 1 – Concept lattice representing the binary relation given in Table 1

Basis on concept lattices

A **formal context** [4] is a triplet (G,M,I) where G is a set of objects, M, a set of attributes and $I\subseteq G\times M$ is a binary relation and for any $g\in G$ and $m\in M$, $(g,m)\in I$ if the object g has the attribute m. Two derivation operators, both denoted $'$, link objects and attributes:

for $O\subseteq G$ and $A\subseteq M$, $O'=\{a\in A \mid \forall o\in O, (o,a)\in I\}$, $A'=\{o\in O \mid \forall a\in A, (o,a)\in I\}$. The compound operators $''$ are closure operators over 2^G and 2^M.

A **concept** c is a pair (O'',O') where $O\subseteq G$. $O''\subseteq G$ is called the **extent** of c and $O'\subseteq M$ the **intent** of c. Both intent and extent are closed sets which intuitively means that the extent of c is the exact set of objects which share all attributes in the intent and no other attribute, and dually between the class of attributes versus objects. The set C of concepts is partially ordered: for any $c_1=(O_1,A_1)$ and $c_2=(O_2,A_2)$, $c_1\leq c_2 \Leftrightarrow O_1\subseteq O_2 \Leftrightarrow A_1\supseteq A_2$. The structure $L(C,\leq)$ defines a lattice, called **concept lattice**.

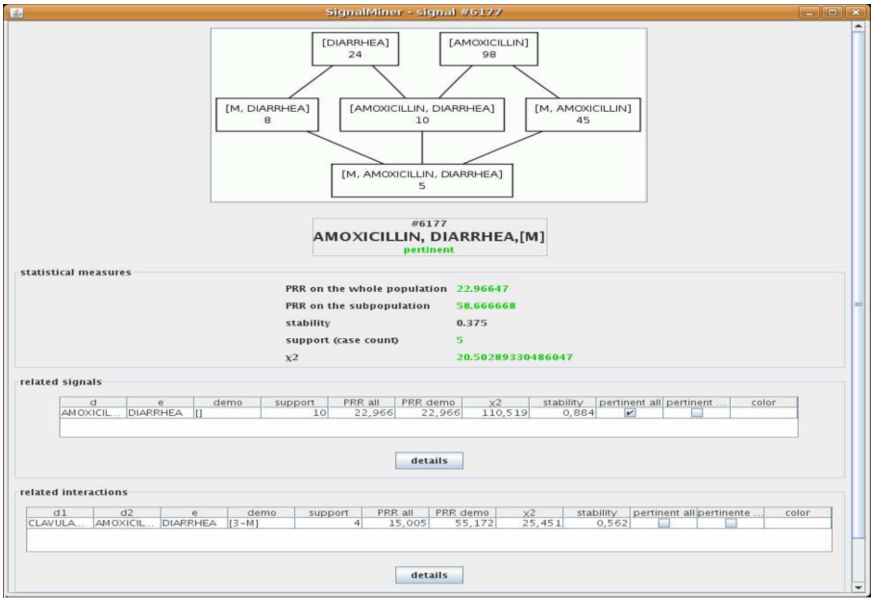
Applied to pharmacovigilance, objects are cases and attributes are drugs $(d_1 \ldots d_5)$, AE $(e_1 \ldots e_{10})$, and demographic attribute such as gender (M,F) and age bracket $(<18,\ldots,>60)$ of the patient (see Table 1). In the resulting concept lattice, shown in Figure 1, concepts are represented by boxes in which the upper (resp. lower) part contains the extent (resp. intent).

A reduced labeling scheme is used so that each object/attribute appears only once in the lattice. An attribute (resp. object) label appears in the highest (resp. lowest) concept that contains it in its intent (resp. extent). A concept labeled with an attribute a is called the attribute-concept of a denoted $\mu(a)$: for instance, $c_1=\mu(e_2)=\mu(e_8)$. Therefore, the intent of a concept is made of all attributes whose attribute-concepts can be reached from the concept on an upward-heading path while extent is recovered in a dual way. For example, considering the concept c_8, its intent contains all the intent labels of its ancestors $\{e_2,e_8,F,e_1\}$, and its extent all the extent labels of its successors $\{\#3,\#4\}$.

The lattice as a synthetic view of the database

The concept lattice gives insights into the case database. We illustrate this point by few examples: $c_8\leq c_1$ since $\{\#3,\#4\}\subseteq\{\#2,\#3,\#4\}$ and $\{e_2,e_8,F,e_1\}\supseteq\{e_2,e_8\}$. This means that among the cases containing adverse effects $\{e_2,e_8\}$, some of them (but not all) are women (F) who also suffer from e_1. By definition, the intent of a formal concept is a closed itemset[1] and the lattice contains all possible closed itemsets as intents. Thus $\{e_2,e_8,F,e_1\}$ (intent of c_8) is a closed itemset but $\{e_2,e_8,F\}$ is not closed as there is no concept with this exact intent. This means that there is no case

[1] Here, the data mining term *itemset* denotes a set of attributes.



SignalMiner - signal #6177

#6177
AMOXICILLIN, DIARRHEA,[M]
pertinent

statistical measures

PRR on the whole population	22.96647
PRR on the subpopulation	58.666668
stability	0.375
support (case count)	5
x2	20.50289330486047

related signals

d	e	demo	support	PRR all	PRR demo	x2	stability	pertinent all	pertinent ...	color
AMOXICIL...	DIARRHEA	[]	10	22,966	22,966	110,519	0,884	☑	□	

details

related interactions

d1	d2	e	demo	support	PRR all	PRR demo	x2	stability	pertinent all	pertinente ...	color
CLAVULA...	AMOXICIL...	DIARRHEA	[3-M]	4	15,005	55,172	25,451	0,562	□	□	

details

Figure 2 – graphical user interface showing the (amoxicillin, diarrhea, M) signal

in the database where a woman suffers from e_2 and e_8 without suffering from e_1. The concept c_9 with $\{e_2,e_8,d_3,d_2\}$ as intent shows that every patient who took d_3 also took d_2 and suffers from e_2 and e_8, since $c_9=\mu(d_3)$.

Thus, through the relations existing between labeled concepts, the lattice reveals correlations between attributes: for instance "d_3 is always taken with d_2", "d_1 is only taken by men", "all and only elderly suffer from e_5"...

A search space for candidate signals and interactions

The lattice is used to identify necessary conditions for both signals (1 drug, 1 AE) and interactions (2 drugs, 1 AE). A necessary condition for a signal to occur is when the intent of a concept contains a pattern (d,e), i.e. it contains exactly one drug and one AE. Such a concept is called a "signal-concept". Note that the intent of a signal-concept may be $\{d,e\}$ or $\{d,e,X\}$ where X is a set of demographic attributes. We note signal-concepts c_{de} or c_{deX}. Similarly, the intent of an "interaction-concept" contains the pattern (d_i,d_j,e). Let us call c_{ij} (or c_{ijX}) an "interaction-concept" with the pattern (d_i,d_j,e). If its related signal concepts c_i with the pattern (d_i,e) and c_j with the pattern (d_j,e) exist in the lattice, by construction, then $c_i \leq c_{ij}$ and $c_j \leq c_{ij}$. The absence of c_i means that d_i and e never appears together without d_j. Thus the lattice defines the search space for candidate signals and interactions. It also links interactions to their related signals.

Statistical measures are then computed for each signal-concept c_{de} or c_{deX} in order to evaluate its pertinence with respect to the British Medicines and Healthcares products Regulatory Agency (MHRA) interestingness criteria [2]. We adopted three criteria for raising hypotheses regarding signals: number of cases ≥ 3, $\chi^2 \geq 4$, and PRR ≥ 2. If the three criteria are successful, the signal-concept c_{de} (resp. c_{deX}) generates a **potential signal** (d,e) (resp. (d,e,X))

For the first criterion, the number of cases is actually the extent's cardinal of the concept. The lattice contains all the information needed to compute the contingency table and therefore the two later measures [5].

Considering a signal-concept c_{de} without demographic attribute, the PRR is computed as follows:

$$PRR(d,e)=\frac{P(e|d)}{P(e|\overline{d})}=\frac{P(de)P(\overline{d})}{P(d)P(\overline{d}e)}=\frac{|de|\cdot|\overline{d}|}{|d|\cdot|\overline{d}e|} \qquad (1)$$

where $|a|$ denotes the number of cases that has attribute a.

Considering a signal-concept c_{deX} with a set of demographic attributes X, the PRR is computed as follows:

$$PRR(d,e,X)=\frac{|(de)_X|\cdot|\overline{d_X}|}{|d_X|\cdot|(\overline{d}e)_X|} \qquad (2)$$

Formula (2) provides a PRR that takes demographic attributes into account by restricting the scope to the concerned population. Hence, $|(de)_X|$ denotes the number of patients in the X subpopulation that took d and suffer from e.

An interaction-concept with the pattern (d_i,d_j,e) becomes a **potential interaction** if it satisfies the three following criteria[2]: number of cases ≥ 3, PRR ≥2 and the interaction's PRR

[2] The PRR value of an interaction (d_1,d_2,e) is computed as follows: $PRR(d_1,d_2,e) = P(e|d_1d_2) / P(e|\ not(d_1d_2))$

has to be greater or equal to each PRR of the two related signals if the signal-concepts of these signals exist in the lattice. The later criterion means that an interaction must "override" its related signals. When it is the case, its related signals are removed from the set of potential signals.

However, related signal-concepts may not exist in the lattice. We proved that it is not worth computing the PRR for non-closed signals (d_i, e) since it is always less or equal to the PRR of the interaction (d_i, d_j, e).

Navigation through results

The user interface makes the most of all the previous observations. Experts can see what kind of relation exists between drug(s) and AE, for instance "all patients that took drug d_i suffer from e". They may access to a set of potential signals and interactions, along with their statistical measures. Each signal and interaction is linked with its concept in the lattice and a subpart of the lattice can be visualized to help experts in their interpretation task.

Figure 2 shows the user interface illustrating a signal d,e,X where d=amoxicillin, e=diarrhea, X={male}. A subpart of the lattice is shown, which contains the concepts c_{deX}, µ(d), µ(e), and all concepts on the paths from c_{deX} to µ(d) and µ(e), here c_{de}, c_{dX}, and c_{eX}.

Concepts are labeled with their intents and the number of objects in extent. Through this graph, experts can observe the distribution of cases in the database: 24 patients suffer from diarrhea (µ(e)), 98 took amoxicillin (µ(d)), 10 took amoxicillin and suffer from diarrhea (c_{de}), and among them 5 are men (c_{deX}).

Then, experts can compare PRR values for both signals c_{de} (denoted "PRR on the whole population") and c_{deX} ("PRR on the subpopulation") and understand why, in this exemple, PRR(d,e,X)>PRR(d,e). It can be seen that 5 men took amoxicillin among the 8 men that suffer from diarrhea; *i.e.* almost all of them. Unlikely, only 10 people took amoxicillin among 24 people who suffer from diarrhea, *i.e.* less that half of them. Thus, the demographic attribute M makes the signal stronger according to PRR values.

In addition, it is possible to compare the strengths on two subpopulations. The same signal on the female population shows a lower PRR (13.38). Hence, we have |(diarrhea, F)|=16 and |(amoxicillin, diarrhea, F)|=5. So the weight of amoxicillin takers within women suffering from diarrhea is lower compared to men, and compared to the whole population. Thus, examining the lattice allows experts to understand why a signal is stronger on given population.

A potential noise detection method

Trimethoprim and sulfamethoxazole come together in marketed drugs, thus a unique concept µ(trimethoprim) = µ(sulfamethoxazole) should exist in the lattice. It is not the case in Figure 3 since only one case has been badly filled in the database. The stability ratio [6] of a concept can capture such a situation. It quantifies the ability of the concept to remain existent after deletion of objects in its extent. Here a low stability can be used to identify concepts in the lattice resulting from noise in the database.

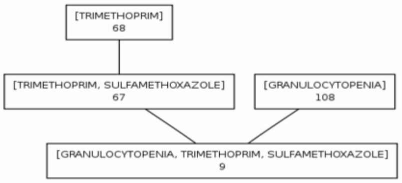

[TRIMETHOPRIM]
68
[TRIMETHOPRIM, SULFAMETHOXAZOLE]
67
[GRANULOCYTOPENIA]
108
[GRANULOCYTOPENIA, TRIMETHOPRIM, SULFAMETHOXAZOLE]
9

Figure 3 – an interaction example containing noise

Results

We applied our method on a subset of the French national SRS database. This subset contains 3249 cases, 976 drugs, 573 AE, and demographic attributes such as gender and age, divided in 3 brackets (<18, 18-60, >60). The resulting lattice contains 13178 concepts, among which 6788 contains at least 3 cases in the extent. The 2812 candidate signal-concepts led to 565 potential signals and the 836 candidate interaction-concepts to 102 potential interactions. Note that the exhaustive search performed by existing method would generate more than 500,000 candidate signals and more than 270,000,000 candidate interactions.

In the worst case, the lattice contains 2^n concepts where n is the minimum of the number of objects and the number of attributes. In practice and especially with SRS databases, the number of closed itemsets is much lower than 2^n and even lower than the number of candidates for exhaustive search.

Table 2 shows the distribution of the 565 signals by pattern. Interestingly, we observe that signals can be divided in four distinct categories, depending on the weight of the demographic attributes. Only 29% of the signals have the pattern (d,e), the rest of the signals have at least one (49%) or two demographic attributes (22%). The validation of such attributes *a posteriori* by manual review of all signals shows a good relevance. In the majority of cases, the demographics 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, aged 18 to 60.

Secondary, all the signals were classified into 5 categories (see Table 3). Categories (1),(2) contain true positives, (3),(4) false positives and (5) unknown potential signals. Table 3 shows that our method generates few false positives that are discussed in the next section. 27 signals were classified as unknown, i.e. not reported in the literature, but interesting enough for investigation by experts.

Table 2 - distribution of the 565 potential signals by pattern

signal pattern	count (%)	example
drug, effect	160 (29%)	cefazolin, thrombocytopenia
drug, effect, gender	132 (23%)	furosemide, gynecomastia, male
drug, effect, age	148 (26%)	abciximab, thrombocytopenia, > 60
drug, effect, gender, age	125 (22%)	levofloxacin, mental confusion, female, > 60

Table 3- Classification of the 565 potential signals

category	signals
(1) known (in reference documents)	502 (89%)
(2) known (in a similar form)	24 (4%)
(3) the AE is the origin of the medication	3 (1%)
(4) due to concomitant drug	9 (2%)
(5) unknown potential signal	27 (5%)

Discussion and perspectives

In this section we discuss the efficiency of our qualitative approach, especially for handling demographic attributes. Then, we present future directions about preventing false positives.

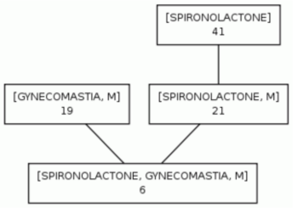


Figure 4 – a signal containing a masculine AE

The lattice helps experts in interpreting PRR variations depending on demographic attributes as shown before for (amoxicillin, diarrhea). Moreover, in some situations where an AE is specific to a population, *e.g.* gynecomastia (see Figure 4), handling demographic attributes in the PRR computation is the only way to obtain a meaningful measure.

False positives (contained in categories (3) and (4)) are common in signal detection. In the following, we give directions for preventing false positives from category (4). The signal (hydrochlorothiazide, cough) is detected because these drug and AE 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 less often than the association (hydrochlorothiazide, cough). Therefore, only this later signal is detected. A solution would be to introduce drug therapeutic families, such as ACE, as attributes, with (o,ACE) ∈ 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 weak to be detected.

In this paper, we have presented a novel automated signal detection method that focuses on the qualitative aspects of the extracted signals. The pivot structure is a concept lattice that allows experts to identify unexpected situations in the case database, and provides information to the experts about why each signal has been detected. Besides our symbolic approach, we have implemented disproportionality measures which are commonly accepted in pharmacovigilance. Our first results based on an extract of the French database are very encouraging: our method has a very good relevance and the signal pattern includes demographic attributes. Further research will focus on (i) improvements for preventing false positives (ii) the scalability of our approach and its efficiency on a bigger database.

References

[1] European Commission, The EU pharmacovigilance system, 2009. http://ec.europa.eu

[2] Hauben M., Madigan D., Gerrits CM., Walsh L., and Van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opinion on Drug Safety, 4(5), Ashley, 2005, pp. 929-948.

[3] Lilienfeld DE. A challenge to the data miners. Pharmacoepidemiology and drug safety, 13(2), John Wiley & Sons, 2004, pp. 881-884.

[4] Ganter B., and Wille R. Formal concept analysis: mathematical foundations. Springer, 1999.

[5] Pogel A., and Ozonoff D. Contingency structures and concept analysis. In proc. of the 6th Int Conf on Formal Concept Analysis, LNCS 4933, Springer, 2008, pp. 305-320.

[6] Kuznetsov S., Obdiekov S., and Roth C. Reducing the representation complexity of lattice-based taxonomies. In proc of the 15th International Conference on Conceptual Structures (ICCS), LCAI 4604, Springer, 2007, pp. 241-254.

Address for correspondence

Jean Villerd, jean.villerd@loria.fr, Loria – INRIA Nancy Grand Est, BP 239, 54506 Vandœuvre-lès-Nancy, France.