Discovering ADE associations from EHRs using pattern structures and ontologies

Gabin Personeni, Marie-Dominique Devignes, Michel Dumontier, Malika Smaïl-Tabbone, Adrien Coulet

Discovering ADE associations from EHRs using pattern structures and ontologies

Gabin Personeni [a], *Marie-Dominique Devignes* [a], *Michel Dumontier* [b], *Malika Smaïl-Tabbone*[a], *Adrien Coulet* [a]

[a] *LORIA (CNRS, Inria NGE, Université de Lorraine), Campus scientifique, Vandœuvre-lès-Nancy, F-54506, France*
[b] *Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305-5479, United States*

ABSTRACT

Patient Electronic Health Records (EHRs) constitute an essential resource for studying Adverse Drug Events (ADEs). We explore an original approach to identify frequently associated ADEs in subgroups of patients. Because ADEs have complex manifestations, we use formal concept analysis and its pattern structures, a mathematical framework that allows generalization, while taking into account domain knowledge formalized in medical ontologies. Results obtained with three different settings show that this approach is flexible and allows extraction of association rules at various levels of generalization.

1 INTRODUCTION

Adverse Drug Events (ADEs) occurs unevenly in different groups of patients. Their causes are multiple: genetic, metabolic, interactions with other substances, etc. Electronic Health Records (EHRs) have been successfully used to detect ADEs (LePendu, et al., 2013). We further hypothesize that mining EHRs may reveal that subgroups of patients sensitive to some drugs are also sensitive to others. We propose a method to identify frequently associated ADEs in these subgroups. Because the way ADEs manifest and are reported is complex and highly variable, we use an extension of formal concept analysis called pattern structures (Ganter & Kuznetsov, 2001) in combination with ontologies to enable generalization. We experimented on a dataset of EHRs from patients diagnosed with Systemic Lupus Erythematosus (SLE), a severe autoimmune disease. Such patients frequently experience ADEs as they often take multiple and diverse drugs indicated for SLE or derived pathologies (Vasudevan & Ginzler, 2009). The SLE EHRs were extracted from STRIDE, an EHR data warehouse of Stanford Hospital and Clinics (Lowe, et al., 2009).

2 MATERIAL AND METHODS

2.1 Data Corpus

Our data corpus is a set of 6,869 anonymized EHRs of patients diagnosed with SLE. It documents about 451,000 hospital visits with their relative dates, diagnoses encoded as a ICD9-CM phenotype codes (International Classification of Diseases, Clinical Modification) and drug prescriptions as a list of their ingredients, represented by their RxNorm identifiers. We use the term "drug" to denote an active ingredient rather than a commercial drug. As we aim to mine frequently co-occurring ADEs, we first need to identify such events in EHRs, and then select patients with at least two ADEs.

We first establish a list of ADE candidates, considering each patient EHR separately. From each two consecutive visits in the EHR, we extract the set of drug D_i prescribed during the first visit and the diagnoses P_i reported during the second. The interval between the two consecutive visits must be less than 14 days, as it is reasonable to think that a side effect should be observed in such a time period after prescription. Moreover, Table 1 shows that increasing that interval does not significantly increase the number of patients in our corpus.

Table 1. Number of patients with at least 2 selected ADEs and number of ADEs for these patients, for different maximum inter-visit interval in days.

Interval (days)	1	2	6	10	**14**	18	22	26	30
|Patients|	434	461	498	526	**548**	555	558	564	576
|ADEs|	2,396	2,587	2,902	3,110	**3,286**	3,388	3,454	3,501	3,621

An ADE candidate C_i is thus a couple of sets $C_i = (D_i, P_i)$. We retained in P_i only phenotypes reported as a side effect for at least one drug of D_i in the SIDER 4.1 database of drug indications and side effects (Kuhn, et al., 2016). We then remove candidates where P_i is empty. Furthermore we remove an ADE candidate (D_1, P_1) if there exists for the same patient another ADE candidate (D_2, P_2) such that $D_1 \subseteq D_2$: indeed, reiterated prescriptions of drugs may indicate that they are safe for this patient.

Table 2. Example of a corpus containing 3 patients and 6 ADEs.

Patient	ADEs
P1	$(\{\text{prednisone}\}, \{\text{ICD 599.8}\})$
	$(\{\text{acetaminophen}\}, \{\text{ICD 599.9}\})$
P2	$(\{\text{prednisone}\}, \{\text{ICD 599.8}\})$
	$(\{\text{prednisone}\}, \{\text{ICD 719.4}\})$
P3	$(\{\text{prednisone, acetaminophen}\}, \{\text{ICD 599.9}\})$
	$(\{\text{acetaminophen}\}, \{\text{ICD 719.4}\})$

After filtering, we obtain a total of 3,286 ADEs from 548 patients presenting at least 2 ADEs. In Table 2, we present examples of ADEs that could be extracted from the EHRs, that will serve as a running example.

2.2 Formal Concept Analysis and Pattern Structures

Formal Concept Analysis (FCA) (Ganter & Wille, 1999) is a mathematical framework for organizing a dataset as a concept lattice, *i.e.*, a hierarchical structure in which a concept represents a set of objects sharing a set of properties. It offers facilities for data mining of association rules.

In classical FCA, a dataset is a set of objects, where each object is described by a set of binary attributes. Pattern structures generalize FCA so that it can be performed on a set of objects with descriptions of any nature, such as sets, graphs, intervals, annotations from ontologies (Ganter & Kuznetsov, 2001; Coulet, et al., 2013). A pattern structure is a triple $(G, (\mathcal{D}, \sqcap), \delta)$, where:

- G is a set of objects, $\mathcal{D}$ is a set of descriptions, δ is a function that maps objects to their descriptions.
- $\sqcap$ is the *meet operator* defining a partial order $\leq_\sqcap$ on elements of $\mathcal{D}$, such that, $X \sqcap Y$ is the most specific description that is more general than both X and Y. Consequently, $X \leq_\sqcap Y$ denotes that Y is more specific than X, is by definition equivalent to $X \sqcap Y = X$. Generalization on object descriptions is performed through the use of the meet operator. This is illustrated for our particular operators in Section 3.

In pattern structures, the derivation operator $.^\square$defines a Galois connection between sets of objects and descriptions, as follows:

$$A^\square = \sqcap_{g \in A}\, \delta(g) \quad \text{for a set of objects } A$$
$$d^\square = \{g \in G \mid d \leq_\sqcap \delta(g)\} \quad \text{for a description } d$$

A pattern concept is a pair (A, d) verifying $A^\square = d$ and $d^\square = A$.

In our case study, G is the set of patients that are related through δ to the description of their adverse drug events from $\mathcal{D}$. Section 3 describes different experiments using pattern structures, each providing their own definition of the triple $(G, (\mathcal{D}, \sqcap), \delta)$.

2.3 Medical Ontologies

We use two medical ontologies, considering only their class hierarchy, to uncover more general phenotype and drug descriptions: ICD9-CM that describes classes of phenotypes and the Anatomical Therapeutic Chemical Classification System (ATC) that describes classes of drugs. We use only the three most specific levels of ATC: pharmacological subgroups, chemical subgroups and chemical substances.

3 EXPERIMENTS

In this Section, we describe three experiments using pattern structures to extract association rules between ADEs. Each experiment defines a different representation of patient ADEs, making increasing use of ontologies.

3.1 First Pattern Structure Experiment

We define here the pattern structure $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$: objects of the pattern structures are patients, and a patient description of $\mathcal{D}_1$ is a vector of sub-descriptions, with first-level ICD classes as dimensions. Each sub-description is a set of drug prescriptions, *i.e.*, a set of sets of drugs. For instance, considering only the two ICD classes in Table 3:

$$\delta_{1,\,\text{ICD 580-629}}(\text{P1}) = \{\{\text{prednisone}\}, \{\text{acetaminophen}\}\}$$
$$\delta_{1,\,\text{ICD 710-739}}(\text{P1}) = \emptyset$$

Table 3. Example of representation of patient ADEs for $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, with two first-level ICD classes: diseases of the genitourinary system (580-629), and of the musculoskeletal system and connective tissue (710-739).

	ICD 580-629	ICD 710-739
Patient P1	$\{\{\text{prednisone}\}, \{\text{acetaminophen}\}\}$	$\emptyset$
Patient P2	$\{\{\text{prednisone}\}\}$	$\{\{\text{prednisone}\}\}$
Patient P3	$\{\{\text{prednisone, acetaminophen}\}\}$	$\{\{\text{acetaminophen}\}\}$

Sub-descriptions are associated to a first-level ICD class to represent ADEs: the patient presents a phenotype of that class after taking a prescription in that sub-description. We define sub-descriptions as sets of prescriptions, where none of the prescriptions are comparable to each other by the partial order $\subseteq$. We then define the meet operator $\sqcap_1$, such that, for every pair of descriptions (X, Y) of $\mathcal{D}_1$:

$$X \sqcap_1 Y = \max(\subseteq, \{x \cap y \mid (x, y) \in X \times Y\})$$

where $\max(\leq_i, S)$ is the unique subset of maximal elements of a set S given any partial order $\leq_i$. Formally, $\max(\leq_i, S) = \{s \mid \nexists x.\ (s \leq_i x)\}$. In the present case, it retains only the most specific drug prescriptions in the description. For instance, given four drugs $d_1$through d_4:

$$\begin{aligned}
&\{\{d_1, d_2, d_3\}\} \sqcap_1 \{\{d_1, d_2\}, \{d_2, d_4\}\} \\
&= \max\big(\subseteq, \{\{d_1, d_2, d_3\} \cap \{d_1, d_2\}, \{d_1, d_2, d_3\} \cap \{d_2, d_4\}\}\big) \\
&= \max\big(\subseteq, \{\{d_1, d_2\}, \{d_2\}\}\big) \\
&= \{\{d_1, d_2\}\}
\end{aligned}$$

We only retain $\{d_1, d_2\}$ since $\{d_2\} \subseteq \{d_1, d_2\}$ and $\{d_1, d_2\}$ is the only $\subseteq$-maximal element. Indeed, the semantic of $\{d_2\}$ – a prescription that contains the drug d_2 – is more general than the semantic of $\{d_1, d_2\}$ – a prescription that contains both the drugs d_1 and d_2.

Given that each patient has a description for each first-level ICD class, the meet operator defined for a single description can be applied to a vector of sub-descriptions as follows:

$$\begin{aligned}&\delta_1(\mathrm{P1}) \sqcap_1 \delta_1(\mathrm{P2})\\ &= \langle \delta_{1,1}(\mathrm{P1}), \dots, \delta_{1,n}(\mathrm{P1})\rangle \sqcap_1 \langle \delta_{1,1}(\mathrm{P2}), \dots, \delta_{1,n}(\mathrm{P2})\rangle\\ &= \langle \delta_{1,1}(\mathrm{P1}) \sqcap_1 \delta_{1,1}(\mathrm{P2}), \dots, \delta_{1,n}(\mathrm{P1}) \sqcap_1 \delta_{1,n}(\mathrm{P2})\rangle\end{aligned}$$

Figure 1 shows the semi-lattice associated with this pattern structure and the data in Table 2. This example illustrates that information about the ADEs is quickly lost as we generalize.

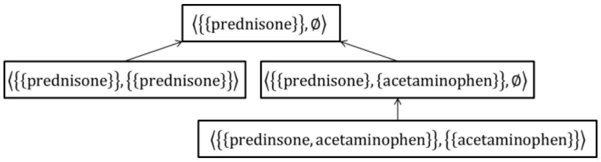
⟨{{prednisone}}, ∅⟩
⟨{{prednisone}}, {{prednisone}}⟩
⟨{{prednisone}, {acetaminophen}}, ∅⟩
⟨{{predinsone, acetaminophen}}, {{acetaminophen}}⟩

Figure 1. Semi-lattice representation of the data in Table 2 using the pattern structure $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, where arrows denote the partial order $\leq_{\sqcap_1}$.

3.2 Extending the Pattern Structure with a Drug Ontology

Using a drug ontology permits to find associations between ADEs related to classes of drugs rather than individual drugs. Thus we extend the pattern structure described previously to take into account a drug ontology: ATC. Each drug is mapped to its ATC class(es), as shown in Table 4.

Table 4. Example of representation of patient ADEs for $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$. Class labels: H02AA03 is desoxycortone, H02AB07 is prednisone, N02BE01 is acetaminophen.

	ICD 580-629	ICD 710-739
Patient P1	$\{\{\mathrm{H02AB07}\}, \{\mathrm{N02BE01}\}\}$	$\emptyset$
Patient P2	$\{\{\mathrm{H02AB07}\}\}$	$\{\{\mathrm{H02AB07}\}\}$
Patient P3	$\{\{\mathrm{H02AB07}, \mathrm{N02BE01}\}\}$	$\{\{\mathrm{N02BE01}\}\}$
Patient P4	$\{\{\mathrm{H02AA03}\}\}$	$\emptyset$

We define this second pattern structure $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$ where descriptions of $\mathcal{D}_2$ are sets of prescriptions with drugs represented as their ATC classes. In order to compare sets of classes from an ontology $\mathcal{O}$, we define an intermediate meet operator $\sqcap_{\mathcal{O}}$, for x and y any two sets of classes of $\mathcal{D}$:

$$x \sqcap_{\mathcal{O}} y = \max\left(\sqsubseteq, \left\{LCA\left(c_x, c_y\right) \mid \left(c_x, c_y\right) \in x \times y\right\}\right)$$

where $LCA\left(c_x, c_y\right)$ is the least common ancestor of c_x and c_y in $\mathcal{O}$, and $\sqsubseteq$ is the ordering defined by the class hierarchy of $\mathcal{O}$. Thus $\max(\sqsubseteq, S)$ is the subset of most specific ontology classes in S, and $x \sqcap_{\mathcal{O}} y$ is the set of most specific ancestors of classes in x and y. From $\sqcap_{\mathcal{O}}$ we define the partial order $\leq_{\mathcal{O}}$ which compares the semantics of a set of ontology classes, such that $x \leq_{\mathcal{O}} y \Leftrightarrow x \sqcap_{\mathcal{O}} y = x$ and $x \leq_{\mathcal{O}} y$ denotes that y is a more specific set of ontology classes than x .

We then define the meet operator $\sqcap_2$ such that for every pair of descriptions (X, Y) of $\mathcal{D}_2$:

$$X \sqcap_2 Y = \max(\leq_{\mathcal{O}}, \{x \sqcap_{\mathcal{O}} y \mid (x, y) \in X \times Y\})$$

This pattern structure allows generalization of ADEs involving different drugs that share a pharmacological subgroup. For instance:

$$\begin{aligned}&\delta(\mathrm{P1}) \sqcap_2 \delta(\mathrm{P4})\\ &= \left\langle\left\{\{\mathrm{H02AB07}\}, \{\mathrm{N02BE01}\}\right\} \sqcap_2 \left\{\{\mathrm{H02AA03}\}\right\}, \emptyset\right\rangle\\ &= \langle\{\mathrm{H02A}\}, \emptyset\rangle\end{aligned}$$

This vector represents the closest generalization of the descriptions of patients P1 and P4, and can be read as: drugs of the class H02A (corticosteroids for systemic use, plain) are associated with a phenotype in the ICD class 580-629, no drugs associated to the ICD class 710-739.

3.3 Extending the Pattern Structure with a Phenotype Ontology

We define a third pattern structure that permits the use of both ATC and ICD for better specialization of phenotypes compared to the previous experiment. Here, we use only the two most-specific levels of ICD, as the previous experiment already covers the most general level. Table 5 describes the data representation used with this pattern structure. Here, ADEs are represented as vectors $\langle D_i, P_i\rangle$ with two dimensions: the set of drugs D_i associated with the set of phenotypes P_i. A patient description is then a set of such vectors.

Table 5. Example of representation of patient ADEs for $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$

	Description
P1	$\{\ \langle\{\mathrm{H02AB07}\}, \{\mathrm{ICD\ 599.8}\}\rangle, \langle\{\mathrm{N02BE01}\}, \{\mathrm{ICD\ 599.9}\}\rangle\ \}$
P2	$\{\ \langle\{\mathrm{H02AB07}\}, \{\mathrm{ICD\ 599.9}\}\rangle, \langle\{\mathrm{H02AB07}\}, \{\mathrm{ICD\ 719.4}\}\rangle\ \}$
P3	$\{\ \langle\{\mathrm{H02AB07}, \mathrm{N02BE01}\}, \{\mathrm{ICD\ 599.9}\}\rangle, \langle\{\mathrm{N02BE01}\}, \{\mathrm{ICD\ 719.4}\}\rangle\ \}$

We define the pattern structure $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$, where descriptions of $\mathcal{D}_3$ are sets of ADEs. We first define an intermediate meet operator $\sqcap_{ADE}$ on our ADEs representations:

$$\begin{aligned}&v_x \sqcap_{ADE} v_y = \langle x_{ATC}, x_{ICD}\rangle \sqcap_{ADE} \langle y_{ATC}, y_{ICD}\rangle\\ &= \begin{cases}\langle x_{ATC} \sqcap_{O} y_{ATC}, x_{ICD} \sqcap_{O} y_{ICD}\rangle \text{ if both dimensions}\\ \qquad\qquad\qquad\qquad\qquad \text{contains non-root classes,}\\ \langle\emptyset, \emptyset\rangle \text{ otherwise.}\end{cases}\end{aligned}$$

The operator $\sqcap_{ADE}$ applies the ontology meet operator $\sqcap_{\mathcal{O}}$ on both dimensions of the vector representing the ADE, using either ATC or ICD as $\mathcal{O}$. Both dimensions of the resulting vector needs to contain non-root ontology classes for it to constitute a representation of an ADE. If it is not the case, we set it to $\langle \emptyset, \emptyset \rangle$ to ignore that ADE in further generalizations.

We define the meet operator $\sqcap_3$ such that for every pair of descriptions (X, Y) of $\mathcal{D}_3$:

$$X \sqcap_3 Y = \max\left(\leq_{ADE}, \left\{v_x \sqcap_{ADE} v_y \mid \left(v_x, v_y\right) \in X \times Y\right\}\right)$$

Compared to $\sqcap_2$, $\sqcap_3$ adds a level of computation with $\sqcap_{ADE}$ which generalizes ADEs, and applies $\sqcap_O$ to an additional ontology: ICD9-CM.

4 RESULTS AND DISCUSSION

The three experiments result in three concept lattices, from which we extract Association Rules (ARs). An AR is identified between two related concepts in the lattice, with descriptions $\delta(l)$ and $\delta(r)$ such that $\delta(r)$ is more specific than $\delta(l)$. Thus, such an AR comprises a left part $L = \delta(l)$ and a right part $R = \delta(r) \backslash \delta(l)$, and is denoted $L \rightarrow R$. Empirically, we only retain ARs with support (number of patients verifying the rule) at least 5, and confidence (ratio of patients verifying L that also verify R) at least 0.75. Table 6 presents a few statistics about this process in our three experiments.

Table 6. Statistics about the lattice building and AR extraction processes implemented in Java.

Experiment	1	2	3
Lattice size (millions of concepts)	1.9	2.3	2.5
ARs extracted (millions)	5	7	9
ARs retained after filtering	772	1,907	913

As expected, this process generates a large amount of rules, among which ARs serving our goal of identifying associations between ADEs must be identified. We therefore define a filter on ARs as the conjunction of the following conditions. *(i)* The right part R of the AR contains at least one ADE, noted as $\langle D_R, P_R \rangle$ for which there does not exist an ADE in the left part L, $\langle D_L, P_L \rangle$, such that either D_R and D_L are $\leq_O$comparable, or P_R and P_L are $\leq_O$comparable. This condition ensures that the right part of the rule introduces new drugs and phenotypes unrelated to those of the left part, *i.e.*, the association between ADEs is not trivial. *(ii)* As patients in the corpus are treated for Systemic Lupus Erythematosus rules must not include related phenotypes (ICD class 710 and descendants).

Finally, we present an example of AR obtained in the third pattern structure experiment, with support 10 and confidence 0.77:

$$\{\langle\{\text{C08DB01}\}, \{\text{ICD 428.0}\}\rangle\} \rightarrow \{\langle\{\text{A02B}\}, \{\text{ICD 427.31}\}\rangle\}$$

This rule means that 77% of patients that present congestive heart failure (ICD 428.0) after prescription of diltiazem (C08DB01), also present atrial fibrillation (ICD 427.31) after prescription of a drug for peptic ulcer and gastro-esophageal reflux disease (A02B). This rule holds true for 10 patients in our corpus. The complete set of filtered rules for each of the three experiments is available online (`http://www.loria.fr/~gpersone/ade-assoc/`).

In summary, we explore in this paper an approach based on pattern structures to mine of EHRs for commonly associated ADEs. It permits a detailed representation of ADEs, enriched with medical ontologies. This representation could be further enriched with additional properties of drugs and phenotypes, such as drugs targets annotated with Gene Ontology classes.

A large amount of ARs can be extracted from our concept lattices. We automatically selected a subset of these ARs by excluding rules that do not fit the scope of the study. We now need to prioritize these ARs with respect to their importance in terms of cost and risk of the phenotypes present in their right part.

One limitation of this work is the absence of temporal relationships between ADEs. We did not consider that aspect because the order of occurrence of ADEs can vary between patients. However, in cases of interest, this order can be checked in patients EHRs as pattern structures concepts retain patient identifiers as well as their description.

REFERENCES

Coulet, A., Domenach, F., Kaytoue, M., & Napoli, A. (2013). Using pattern structures for analyzing ontology-based annotations of biomedical data. *International Conference on Formal Concept Analysis.* Springer.

Ganter, B., & Kuznetsov, S. O. (2001). Pattern Structures and Their Projections. *9th International Conference on Conceptual Structures*, 129-142.

Ganter, B., & Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations.* Springer-Verlag.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research*.

LePendu, P., Iyer, S. V., Bauer-Mehren, A., Harpaz, R., Mortensen, J. M., Podchiyska, T., et al. (2013). Pharmacovigilance Using Clinical Notes. *Clinical Pharmacology & Therapeutics*, (pp. 547-555).

Lowe, H. J., Ferris, T. A., Hernandez, P. M., & Weber, S. C. (2009). STRIDE–An Integrated Standards-Based Translational Research Informatics Platform. *AMIA Annual Symposium Proceedings*, 391-395.

Vasudevan, A. R., & Ginzler, E. M. (2009). Established and Novel Treatments for Lupus. *The Journal of Musculoskeletal Medicine*.