Automatic Creation and Refinement of the Clusters of Pharmacovigilance Terms

Marie Dupuch
CRC, Université Pierre et Marie Curie - Paris 6; Inserm, UMRS 872, Paris, F-75006, France
marie.dupuch@crc.jussieu.fr

Cédric Bousquet
DSPIM, Université de Saint Etienne, F-42023; Inserm, UMRS 872, Paris, F-75006, France
cedric.bousquet@chu-st-etienne.fr

Natalia Grabar
CNRS UMR 8163 STL; Université Lille 1&3
F-59653 Villeneuve d'Ascq
natalia.grabar@univ-lille3.fr

ABSTRACT

Pharmacovigilance is the activity related to the collection, analysis and prevention of adverse drug reactions (ADRs) induced by drugs or biologics. The detection of adverse drug reactions is performed thanks to statistical algorithms and to groupings of ADR terms. Standardized MedDRA Queries (SMQs) are the groupings which become a standard for assisting the retrieval and evaluation of MedDRA-coded ADR reports all through the world. Currently 84 SMQs have been created manually by experts, while several important safety topics are not yet covered. Dependent on the context of their application, these SMQs show a high degree of sensitivity and often appear to be over-inclusive. For pharmacovigilance experts it represents an important and tedious filtering of data. The objective of this work is to propose an automatic method for assisting the creation of SMQs and also for the refinement of their organization further to the creation of smaller clusters of ADR terms. In this work we propose to exploit the semantic distance and clustering approaches. We perform several experiments and vary several parameters of the method.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; I.5.3 [**Pattern Recognition**]: Clustering; J.3 [**Computer Applications**]: Life and Medical Sciences

General Terms

Applications, Experimentation

Keywords

Pharmacovigilance, MedDRA, semantic distance, clustering of terms

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IHI'12, January 28–30, 2012, Miami, Florida, USA.
Copyright 2012 ACM 978-1-4503-0781-9/12/01 ...$10.00.

1. INTRODUCTION

Pharmacovigilance is the activity related to the collection, analysis and prevention of adverse drug reactions (ADRs) likely to be caused by drugs or biologics. The collection of ADRs is achieved thanks to the case reporting to the pharmacovigilance authorities and also to the pharmaceutical industries by medical doctors or by pharmacists. Before their inclusion in pharmacovigilance databases, the ADRs of these case reports are coded with terms from dedicated terminologies, such as MedDRA (Medical Dictionnary for Drug Regulatory Activities) [5]. The analysis of the collected ADRs is related to the safety surveillance within these databases. It often relies on the identification of signals, that are unexpected relations or not yet well defined relations between a drug and an ADR. Statistical methods are typically used in the analysis process [18, 2], nevertheless, it has been observed that some pairs {*drug*, *adverse reaction*} are not activated, when they should be. The main cause then is that MedDRA is a fine-grained terminology containing over 85,000 terms and that the encoding of the adverse reactions with these terms may have an impact on the signal dissolution [9]. This means that similar and close ADRs may be encoded with different MedDRA terms, in which case, during the analysis of the databases they will remain isolated and the safety risk detection may be under-estimated. For instance, terms such as *Hepatitis infectious*, *Hepatitis infectious mononucleosis* or *Hepatitis viral* are different although they mean close and medically related ADRs. When mining the pharmacovigilance databases, it may be useful first to cluster together semantically and medically close terms and then to exploit these clusters for the satefy surveillance [10].

In that purpose, SMQs (Standardized MedDRA Queries) have been created. At the heart of the SMQs is a precise medical definition of a pathology and the SMQs tend to group the terms associated with this pathology. The SMQs are defined by groups of experts through a manual study of both the MedDRA's structure and the scientific literature [6]. It is a long and meticulous task. Now there are 84 SMQs that cover several important medical conditions, as for instance *Glaucoma*, *Hypertension*, *Cardiomyopathy* or *Retinal disorders*. But several other SMQs are still to be defined.

Evaluation studies of the SMQs have demonstrated that SMQs often present a very high sensibility [20, 24], and tend to be over-inclusive [24]. In such a case, the eval-

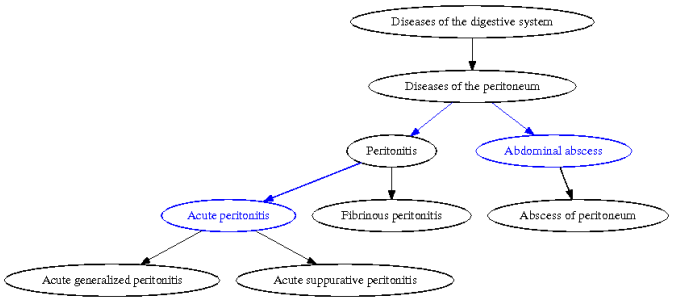

Figure 1: Graph or path-based distance between two terms.

uation of case reports found with the SMQs can be very time-consuming because these reports might lack specificity: their treatment by experts is then very long and tedious. A solution might be the creation of hierarchically structured SMQs, which can be exploited and combined among them to obtain a higher specificity. Among the 84 existing SMQs only 20 are provided with a hierarchical structure.

2. OBJECTIVES

The objective of this work is to explore and to adapt automatic methods which could be used for assisting the building process of the SMQs and also for the refinement of the SMQs' structure.

More precisely, we propose to exploit the semantic distance approaches. Several of these approaches are applied within the tree structures [26, 32, 14, 22], such as terminologies or ontologies, and rely on the number of edges (links) between the two terms in order to compute the semantic distance between these terms. The simplest approach [26], which was the first of its kind, relies on counting edges between terms and on finding the shortest path between them. Thus, on figure 1 we have an excerpt from a terminological graph with nine nodes. When we compute the shortest path between the two blue nodes *Acute peritonitis* and *Abdominal abscess*, we follow the blue path and obtain the shortest distance equal to three edges. In addition to the path length, other criteria may be taken into account: hierarchical depth of terms [30, 33], information content [27], the nearest common parent [15], etc. Besides the computing of the semantic similarity between two terms or words, these approaches have been used in different contexts such as: word-sense disambiguation [30], information retrieval [13, 33], gene anotation [16], terminology enriching and adaptation [31, 8].

In a previous work of our group, the semantic distance was applied to a subset of pharmacovigilance terms [4, 11], and the obtained groupings demonstrated several types of relations: synonyms, antonyms, physiological functions or abnormalities, associated symptoms, abnormal laboratory tests, pathologies and their causes, close anatomical localizations, degrees of severity, and several heterogeneous groupings. None of them could be used as the basis for SMQs creation nor appeared to be close to the content of the SMQs. Another work of our group proposed to create groupings of pharmacovigilance terms on the basis of hierarchical subsumption (terminological reasoning) [12]. These results are compared with 24 SMQs and we will refer to these results in the discussion section. Despite the availability of the semantic distance and terminological reasoning approaches, their application for this kind of task remains an objective hard to reach. In our work, we propose several experiments and tests in order to adapt the semantic distance approaches to the creation of clusters of semantically and medically related pharmacovigilance terms.

Level	Expanded form	Nb Terms
SOC	System Organ Class	26
HLGT	High Level Group Terms	332
HLT	High Level Terms	1,688
PT	Preferred Terms	18,209
LLT	Lowest Level Terms	66,587
Total		86,842

Table 1: Hierarchical levels of MedDRA: number of terms per level.

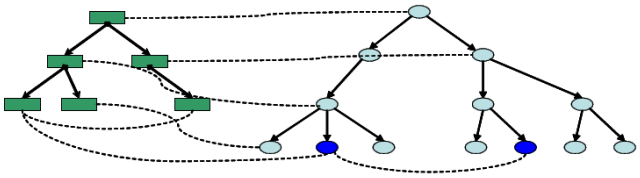

Figure 2: Projection of the MedDRA terms (on the left) towards the SNOMED CT terms (on the right), as illustrated in [1].

3. MATERIAL

The exploited material is specific to the pharmacovigilance area. We exploit terms from the MedDRA terminology, designed for the encoding of adverse drug reactions induced by drugs. It contains a large set of terms (signs and symptoms, diagnostics, therapeutic indications, complementary investigations, medical and surgical procedures, medical, surgical, family and social history). These terms are structured within five hierarchical levels indicated in table 1: *SOC* (*System Organ Class*) terms belong to the highest level, while *LLT* (*Lowest Level Terms*) terms belong to the lowest level. Terms from the *PT* (*Preferred Terms*) level are usually exploited in the pharmacovigilance safety surveillance. Most often, the role of the *LLT* terms is to provide the *PT* terms with synonyms or equivalent terms, although it happens that they have hierarchical relations with *PT* terms [17].

3.1 Ontology ontoEIM

The ontology of adverse drug reactions ontoEIM [1] has been created through the projection of MedDRA on the terminology SNOMED CT [29], as illustrated on figure 2. This projection is performed thanks to the exploitation of the UMLS [23], where an important number of terminologies are already merged and aligned, among which MedDRA and SNOMED CT. Note that the current rate of alignment of the *PT* MedDRA terms with those from SNOMED CT is rather weak: 51.3% (7,629 terms). The projection of MedDRA on SNOMED CT aims at improving the representation of MedDRA terms. The first advantage is that the structuring of MedDRA terms becomes parallel to the structuring

ID	*Names of the hierarchical SMQs*	*Number of*		*PT*	*PT+LLT*
		levels	*s-smq*		
20000074	Adverse pregnancy outcome	2	4	1683	6013
20000118	Biliary disorders	3	11	176	699
20000049	Cardiac arrhythmias	4	12	131	662
20000060	Cerebrovascular disorders	3	5	198	861
20000035	Depression and suicide/self-injury	2	2	137	1028
20000100	Drug abuse, dependence and withdrawal	2	2	42	568
20000081	Embolic and thrombotic events	2	3	277	1048
20000095	Extrapyramidal syndrome	2	4	92	588
20000137	Gastrointestinal nonspecific inflammation	2	3	138	835
20000103	Gastrointestinal perforation, ulceration	2	5	309	1760
20000027	Haematopoietic cytopenias	2	4	119	452
20000038	Haemorrhages	2	2	422	2113
20000170	Hearing and vestibular disorders	2	2	100	486
20000005	Hepatic disorders	4	13	333	1201
20000043	Ischaemic heart disease	2	2	107	585
20000090	Malignancies	2	4	1839	8036
20000109	Oropharyngeal disorders	2	5	250	1104
20000085	Premalignant disorders	2	5	248	821
20000066	Shock	2	6	179	961
20000159	Thyroid dysfunction	2	2	160	701

Table 2: SMQs with hierarchical organization of their terms. We indicate the number of hierarchical levels and of sub-SMQs, and also the number of PT terms and PT and LLT terms.

in SNOMED CT, which makes it more fine-grained [1]: the SNOMED CT-like hierarchy is constructed and new terms are added to fill in the intermediate levels absent among MedDRA terms. The maximal number of the hierarchical levels within the ontoEIM resource can reach up to 14, while only five levels are provided in MedDRA. This improvement makes the application of the semantic distance and similarity measures a well-founded solution. Another advantage is that the MedDRA terms receive formal definitions. Thus, terms can be defined on up to four axes from SNOMED, exemplified here through the term *Arsenical keratosis*:

- *Morphology* (type of abnormality): *Squamous cell neoplasm, Morphologically abnormal structure*;
- *Topography* (anatomical localization): *Skin structure, Structure of skin and or surface epithelium*;
- *Causality* (agent or cause of the abnormality): *Arsenic AND OR arsenic compound*;
- *Expression* (manifestation of the abnormality in the organism): *Abnormal keratinization.*

The names of the formal definition axes (*Morphology, Topography*, etc.) historically correspond to the names of the semantic hierarchies of the Snomed International [7], but the definitions themselves have been extracted from the SNOMED CT resource. Note that the formal definitions are not complete either: only 12 terms receive formal definitions with these four axes and 435 terms are defined with three of the four axes. 2,846 terms have definitions with two axes, and 1,695 more with only one axis. On the one hand, this is due to the fact that the projection of MedDRA terms is not complete, or that there are missing relations, often with morphology terms. On the other hand, these four elements are not relevant for every term and their absence is not always wrong. Despite the shortcomings of this material, ontoEIM (MedDRA terms, their structuring and formal definitions) is our main material exploited for the creation of clusters of adverse drug reactions.

3.2 Standardized MedDRA Queries

Among the 84 existing SMQs, we exploit mainly the 20 SMQs which have a hierarchical structure. In table 2, we indicate the names of these SMQs as well as the number of hierarchical levels, the number of their sub-SMQs and the number of PT and PT+LLT terms they contain. These hierarchical SMQs are structured in different ways. For instance, some SMQs have several hierarchical levels: *Cardiac arrhythmias* and *Hepatic disorders* are divided into up to four levels of sub-SMQs. Consequently, they have a large number of sub-SMQs: 12 and 13 respectively. Although, the majority of the hierarchical SMQs has only two hierarchical levels, and the number of their sub-SMQs varies from two to six. Let us show some examples on how the hierarchical SMQs may be organized. The SMQ 20000060 *Cerebrovascular disorders* has three hierarchical levels and five sub-SMQs (in brackets we indicate the numbers of *PT* terms at a given level):

- *Cerebrovascular disorders* (198)
  - *Central nervous system haemorrhages and cerebrovascular conditions* (30)
    - *Ischaemic cerebrovascular conditions* (67)
    - *Haemorrhagic cerebrovascular conditions* (35)
    - *Conditions associated with central nervous system haemorrhages and cerebrovascular accidents* (30)
  - *Cerebrovascular disorders, not specified as haemorrhagic or ischaemic* (18)

The ADR terms of this SMQ are categorized either under the sub-SMQs or directly under the global SMQ. As for the SMQ 20000038 *Haemorrhages*, it has only two hierarchical levels and only two sub-SMQs:

- *Haemorrhages* (422)
  - *Haemorrhage terms (excl laboratory terms)* (331)
  - *Haemorrhage laboratory terms* (91)

All the ADR terms are categorized within the sub-SMQs: no direct dependencies of terms exists with the global SMQ.

We exploit the 20 SMQs and their 92 sub-SMQs (2010 version) as the gold standard for the evaluation of the clusters of terms we generate with our approach. The evaluation is thus performed at two levels: at the level of the whole SMQs and at the level of their sub-SMQs.

4. CREATION OF CLUSTERS OF THE MEDDRA TERMS AND THEIR REFINEMENT

The proposed method is organized in three main steps: (1) computing of the semantic distance and similarity between MedDRA terms, (2) clustering of the MedDRA terms, (3) and evaluation of the obtained clusters against the SMQs and sub-SMQs. Figure 3 illustrates the steps of the method. For the implementation, we exploit Perl and R[1] languages.

4.1 Computing of the semantic distance and similarity between terms

Semantic distance is computed between the 7,629 *PT* MedDRA terms present in the ontoEIM resource. We exploit only the *PT* terms because they contitute the SMQs, they are used for the coding of the pharmacovigilance case reports worlwide, and if necessary they can bring their *LLT* terms. During this step, we exploit the approaches (one semantic distance and two semantis similarities) to compute the distance between two terms (or terms) *c1* and *c2*:

- the *Rada* approach [26] computes the distance and relies on the detection and computing of the shortest path *sp*, which corresponds to the sum of the edges of this shortest path:

$$sp(c1, c2)$$

- the *LCH* Leacock and Chodorow approach [14] computes the similarity and relies on the shortest path *sp* and on the maximal depth *MAX* found within the terminology (*MAX*=14 within the ontoEIM):

$$-log[\frac{sp(c1, c2)}{2 * MAX}]$$

- the *Zhong* approach [33] computes the distance and relies on the absolute depth *depth* of terms and on their closest common parent *ccp*. The milestone value m is computed first for each term:

$$m(c) = \frac{1}{k^{depth(c)+1}}$$

where c is a term, *depth* its absolute depth within a terminology and $k = 2$ (normalization coefficient). Then, the distance between two terms is computed:

$$2 * m(ccp(c1, c2)) - (m(c1) + m(c2))$$

where *ccp* is the nearest common parent and m milestone values obtained previously.

Semantic distance and similarities are computed between the MedDRA terms but also between the elements of their formal definitions. More precisely, within the formal definitions, we exploit elements provided by two axes: morphology M (type of the abnormality) and topography T (anatomical localization). Very often, these axes are involved in the definition of diagnostics [28] and they are also the most frequent in the ontoEIM resource. As for two other axes (causality C and expression E), as they seldom appear in formal definitions of ontoEIM, we cannot rely on them for the computing of semantic distance and similarity. Formal definitions are exploited in order to improve the semantic representation of terms and in order to make this representation more fine-grained [25]. For the illustration of the approach, let's consider two ADR terms, *Abdominal abscess* and *Pharyngeal abscess* defined as follows:

- *Abdominal abscess*: M = *Abscess morphology*, T = *Abdominal cavity structure*
- *Pharyngeal abscess*: M = *Abscess morphology*, T = *Neck structure*

In the definition of *Pharyngeal abscess*, the anatomical localization is underspecified (*Neck structure*), which actually corresponds to the relations found within the SNOMED CT. Currently we do not complete nor check out the correctness of the formal definitions, although this could be planned for the future. Figure 4 illustrates how the shortest paths *sp* are computed between these two ADR terms and between the elements of their formal definitions (axes T and M). The weight of edges is set to 1 because all relations are of the same kind (hierarchical), and the value of each shortest path corresponds to the sum of weights of all its edges. For this pair of terms we obtain the following values: $sp_{ADR} = 4$, $sp_T = 10$ and $sp_M = 0$. The computing of the semantic distance and similarity is then performed according to the three approaches described above: *Rada*, *LCH* and *Zhong*. The obtained semantic distances or similarities *sd* are then exploited to compute the unique distance between the ADR terms:

$$\frac{\sum_{i \in \{ADR, M, T\}} W_i * sd(c1_i, c2_i)}{\sum_{j \in \{ADR, M, T\}} W_j}$$

where $\{ADR, M, T\}$ respectively correspond to terms meaning the *ADR*, axis Morphology M and axis topography T; $c1$ and $c2$ are two ADR terms; W is the coefficient associated with each of the three terms; and *sd* is the semantic distance or similarity computed on a given axis.

We carry out several experiments and vary several factors:

1. Formal definitions: (1) formal definitions are taken into account and the semantic distance or similarity is computed on three paths, or (2) formal definitions are not taken into account and the semantic distance or similarity is computed on the path of ADRs only;
2. Weights W put on the *ADR* terms and on M and T axes of the formal definitions are set either to 1 or to 2 and all the possible combinations are tested;

[1] http://www.r-project.org

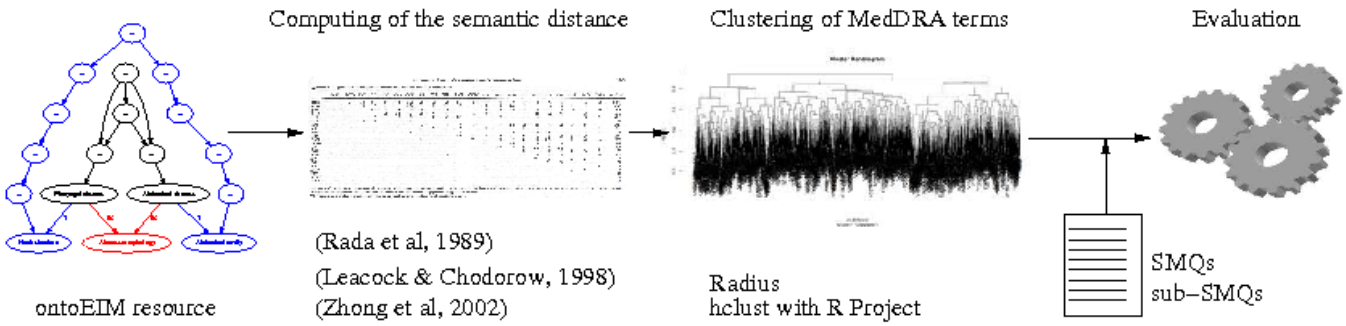

Figure 3: General schema of the method.

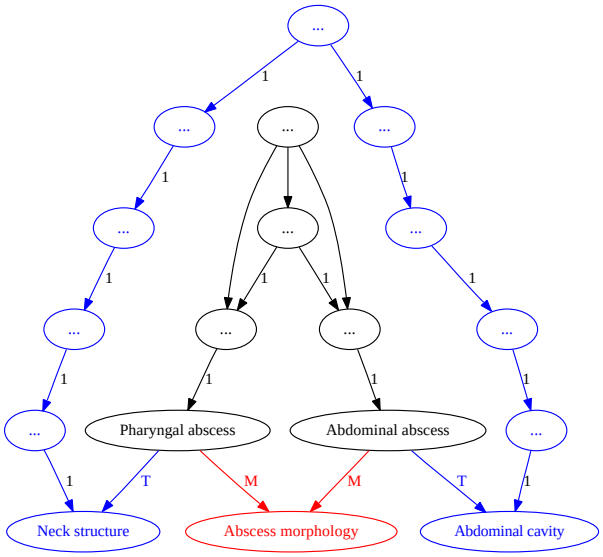

Figure 4: Computing of the shortest paths *sp* between two MedDRA terms (*Abdominal abscess* and *Pharyngeal abscess*) and between the elements of their formal definitions (axis T in blue and axis M in red).

3. Semantic distance and similarity: the three approaches (*Rada*, *LCH* and *Zhong*) are tested.

Further to the application of this method, a semi-matrix 7629*7629 is built. It contains the semantic distances and similarities between the ADR terms, i.e. 7,629 *PT* terms.

4.2 Clustering of terms

Once the distances and similarities are computed, we use them for the creation of clusters of terms. We exploit two approaches for the clustering of ADR terms on the basis of the semantic distances and similarities among them:

- R radius approach, where every ADR term is considered as a possible center of a cluster and its closest terms are clustered together with it. We test several threshold values with the three semantic distance approaches. The obtained clusters are not exclusive and their intersection in not empty.
- *HAC* hierarchical ascendant classification is performed through the *R Project* tools. This method first chooses the best centers for clusters and then builds the hierarchy of terms by progressively merging smaller clusters to obtain the bigger ones. We exploit the function *hclust*. For this function to be applied, we perform several steps. First, the matrix must be converted into the specific format to be read and processed by the *R Project* tools. We then call for a function which reads the matrix and records its content together with the labels of terms:

  read.table("matrix", header=TRUE, sep=";", fill=TRUE)->data
  rownames(data) <- data$PT

  The next step consists in computing the euclidean distance between the 7,629 terms of the matrix:

  d <- dist(data[1:7628,2:7629], method="euclidean")

  We then apply the *hclust* function for the hierarchical clustering of the data and for the creation of a dendrogram with the method *ward*. This method *ward* considers the union of every possible cluster pair at each step, the two clusters whose fusion results in minimum increase in information loss are combined.

  fit <- hclust(d, method="ward")

  Finally, the dendrogram is segmented into n clusters (3,500 in this example):

  cutree(fit,3500) -> fit.id

  The obtained clusters are exclusive and their intersection is empty.

4.3 Evaluation of the generated clusters

The evaluation of the generated clusters is performed thanks to their comparison with the SMQs and sub-SMQs. 20 hierarchical SMQs and their 92 sub-SMQs are exploited as a gold standard in this experiment. A quantitative evaluation is performed with the three classical measures: precision P (percentage of the relevant terms clustered divided by the total number of the clustered terms), recall R (percentage of the relevant terms clustered divided by the number of terms in the corresponding SMQ) and F-measure F (the harmonic mean of P and R). The association between the SMQs and the clusters relies on precision values, because the experts favour the precision. Different precision values have been tested. During this step, we evaluate either one best cluster

	Best			*Merging*		
	P	*R*	*F*	*P*	*R*	*F*
$Rada_r$	78.47	22.86	35.40	38.69	49.16	43.30
$Rada_{hac}$	90.30	15.02	25.75	60.54	30.00	40.11
Lch_r	78.47	22.59	35.08	45.80	45.80	45.80
Lch_{hac}	89.15	15.42	26.29	59.83	31.33	41.12
$Zhong_r$	42.02	16.44	23.63	31.50	25.45	28.15
$Zhong_{hac}$	91.93	15.76	26.90	62.80	30.00	40.60

Table 3: Average precision, recall and f-measure at the level of sub-SMQs. Parameters: without formal definitions (computing of the semantic distance for ADR terms only)

	Best			*Merging*		
	P	*R*	*F*	*P*	*R*	*F*
$Rada_r$	63.45	23.00	33.76	40.64	42.78	41.68
$Rada_{hac}$	90.23	16.67	24.27	61.93	31.89	42.10
Lch_r	75.66	17.61	28.57	51.15	31.49	38.98
Lch_{hac}	85.97	16.91	28.26	62.89	29.56	40.22
$Zhong_r$	40.38	14.36	21.81	31.06	22.58	26.14
$Zhong_{hac}$	91.31	15.72	26.82	63.17	32.10	42.56

Table 4: Average precision, recall and f-measure at the level of sub-SMQs. Parameters: formal definitions (computing of the semantic distance on three axes)

for each SMQ or the n best clusters in which case these clusters are merged. The n value is set automatically according to the SMQs and sub-SMQs and depends on the content of the clusters. It varies between one and 50 clusters merged. As for the size of clusters, it varies between two and 162 terms. A qualitative evaluation is also performed, in which we study the content of the generated clusters and perform a failure analysis.

5. RESULTS

The 7,629 ADR terms from MedDRA have been fully processed through the three semantic distance algorithms and through the two approaches of clustering. With the clusting approaches, we tested several thresholds. Thus, with the R approach the thresholds tested for the semantic distance correspond to the following intervals: two singletons 2 and 3 for *Rada*, [0; 5.059] for *LCH* and [0; 0.49] for *Zhong*. In respect with our data, the best thresholds appear to be: 2 for *Rada*, 4.10 for *LCH* and 0 for *Zhong*. As for the *hclust* clustering, we tested several numbers of classes within the interval [100; 7,000], because the number of ADR terms is 7,629. The best results are obtained with 3,500 classes. Number of terms per class varies from 2 to 14. The results we present and discuss have been obtained with these thresholds: with the Radius approach the thresholds are set to 2 for *Rada*, to 4.10 for *LCH* and to 0 for *Zhong*, while with *hclust* the number of classes is set to 3,500. During the evaluation, the association between the generated clusters and the gold standard is performed on the basis of the percentage of common ADR terms (precision). We tested several precision values within the interval [10; 90]. 50% appears to be the best precision rate with SMQs, while 30% is the best with sub-SMQs. We also varied several other factors when setting up these algorithms. The obtained clusters have been compared with the SMQs and sub-SMQs introduced in the Material section.

The obtained evaluation results are presented in tables 3 and 4 for sub-SMQs, and in tables 5 and 6 for the whole SMQs. Moreover, we distinguish the experiments in which the formal definitions have been exploited (tables 4 and 6) or not (tables 3 and 5). In these tables, for each semantic distance and clustering method, we indicate the average values for the precision, the recall and the f-measure. The first set of the evaluation values is indicated only when the best cluster is exploited (*Best*), the second set of the evaluation values is indicated when n best clusters are grouped together.

	Best			*Merging*		
	P	*R*	*F*	*P*	*R*	*F*
$Rada_r$	99.70	16.85	28.80	63.25	40.65	50.00
$Rada_{hac}$	100.0	5.85	11.05	63.80	18.45	28.62
Lch_r	99.70	16.85	28.80	63.25	40.65	50.00
Lch_{hac}	97.50	6.00	11.13	62.70	18.50	28.60
$Zhong_r$	72.15	5.70	10.56	41.80	30.60	35.33
$Zhong_{hac}$	100.0	6.80	12.80	60.50	19.70	29.66

Table 5: Average precision, recall and f-measure at the level of SMQs. Parameters: without formal definitions (computing of the semantic distance for ADR terms only)

	Best			*Merging*		
	P	*R*	*F*	*P*	*R*	*F*
$Rada_r$	95.05	16.80	28.55	63.25	43.95	51.86
$Rada_{hac}$	100.0	6.40	12.03	64.35	17.60	27.60
Lch_r	93.90	11.30	20.17	57.20	27.25	37.00
Lch_{hac}	97.50	6.45	12.10	65.45	17.75	28.00
$Zhong_r$	66.45	7.10	12.80	37.40	26.95	31.30
$Zhong_{hac}$	100.0	7.95	14.70	55.00	27.35	36.53

Table 6: Average precision, recall and f-measure at the level of SMQs. Parameters: formal definitions (computing of the semantic distance on three axes)

We can observe that globally the semantic distance approach generates the clusters which show a very high precision at both evaluated levels: SMQs and sub-SMQs. When the merging of n best clusters is performed, the recall is improved while the precision decreases. With sub-SMQs, the *hclust* clustering method provides with a better precision than with the Radius clustering. We discuss these different results and observations with more details in the next section.

6. DISCUSSION

6.1 Material

The ontoEIM resource, exploited in this work, is currently built thanks to the projection of MedDRA terms on the Snomed CT. Since this process is performed through the UMLS the alignment between these two terminologies are those already defined in the UMLS. As we focused on in the material section, the alignment is not perfect. This situation has a negative effect on the success of the method we apply for clustering the terms. Our group is currently working on the improvement of the alignment rate. We exploit the existing work on the alignment of the MedDRA terms [3, 21] and also perform additional experiments [19] to improve the current alignment rate. Lexical mapping and semantic categorization methods allow to reach both a high sensitivity and the semantic constraints. Nevertheless, this task is not

finished yet and the full mapping of the MedDRA terms still requires more effort.

6.2 Comparison of the clusters with SMQs and sub-SMQs

The generated clusters have been evaluated with the SMQs and sub-SMQs. The applied method provides a high precision. This means that several small and precise clusters of terms are generated. In tables 3 and 4, at the levels of the sub-SMQs, we can clearly observe that the best cluster always provides with a very high precision and covers up to 23% of the required terms, although it may contain irrelevant ADR terms as well. When the n best clusters are merged, on the one hand the recall is improved because more relevant terms are grouped together, and on the other hand the irrelevant terms are accumulated and this causes the decrease of precision. This result meets the expectations of the pharmacovigilance experts. Indeed, they need clusters of the SMQ terms which are smaller than the SMQs and which, for this reason, will give more precise sets of the pharmacovigilance cases. At the level of SMQs (tables 5 and 6), the precision is still very high with the best clusters, but the recall becomes really low. Indeed, the SMQs are much larger than their sub-SMQs and this yields normally the lesser recall. Up to now, we presented and discussed the average evaluation figures, but there is a variability in results and the situation varies a lot according to SMQs and sub-SMQs. In table 8, we present the results for two hierarchical SMQs. For each of these SMQs, we first indicate the figures for the whole SMQ and then the figures by its sub-SMQs. The performances in retrieving the same PT terms depend on the number of terms within the SMQ of sub-SMQ. We can indeed observe that the performances are higher at the level of the sub-SMQs.

Hence, our results show that the semantic distance methods are suitable for the generation of reduced and precise clusters of ADR terms. In our opinion, these clusters can be exploited in two contexts:

- *Creation of the SMQs.* The clusters can be used for the creation of the sub-SMQs and of the SMQs. In this case, they will help the experts involved in the building of the SMQs and will especially be helpful for browsing the MedDRA terms and for the collection of semantically close ADR terms.

- *Mining the pharmacovigilance databases.* The clusters can also be used for the segmentation of the existing SMQs into several homogeneous subsets. Such clusters can indeed be exploited for the safety surveillance and would reduce the set of pharmacovigilance cases to filter and to analyse.

Our results are encouraging, but a thorough evaluation of the clusters within these contexts is still to be performed. Let's compare our results to those previously obtained in our group in which the terminological reasoning was applied [12]: in this work the sensitivity (or precision) is evaluated and shows the average value of 0.82, while the specificity is not evaluated. In several experiments and settings of our work, the average precision obtained is higher and we also evaluate the recall. Besides the grouping approach exploited, the main difference is related to the material: WHO-ART terms (which alignment reaches up to 85.9%) and subsets of SMQs are exploited in the previous work [12], while we exploit the MedDRA terms and the whole SMQs.

6.3 Factors which influence the performances

We performed several experiments and varied several setting at different steps of the method. We discuss here those which influence the results:

- *Semantic distances.* We applied three semantic distances or similarity algorithms (*Rada*, *LCH* and *Zhong*). It is interesting to observe that the most simple algorithm *Rada*, which relies only on the number of edges, appears to be the most efficient. The *LCH* algorithm provides also good results close to those of the *Rada* algorithm. The common feature between these two algorithms is the shortest path *sp*, which means that the maximum depth *MAX* involved in the *LCH* algorithm has no impact. As for the *Zhong* approach, which also exploits the absolute depth of terms, this criteria does not seem to be relevant for clustering the pharmacovigilance terms. Indeed, in this task, it may be important to cluster terms from lower and also from higher hierarchical levels, while the *Zhong* algorithm favours hierarchically lowest terms.

- *Formal definitions.* An important difference is observed in relation with the exploitation of the formal definitions: when we use only the ADR terms, the performances are always better than when we use these terms with their formal definitions. This situation is due to the incompleteness of the currently available formal definitions. As we reported in the Material section, only a small number of terms are defined on the exploited axes. But a manual evaluation of several incorrect clusters, did show that the formal definitions of the involved terms where not complete. In such cases, the semantic distance measure is favouring only the defined axis within the formal definitions (morphology or topography), which leads to a distorted semantic representation of the terms and to a wrong clustering.

- *Coefficients of axes.* We tested two coefficients, 1 or 2, associated with ADR terms and the axes of formal definitions. The weighting of axes allows indeed to give more importance to one of the aspects of the definitions of ADR terms. All the possible combinations were set up and evaluated. The best results are obtained when the coefficients have the following values: $W_{ADR} = 1$, $W_M = 2$, $W_T = 1$. This result suggests that the morphology axis M is the main factor for the clustering of ADRs because it specifies the kind of morphological abnormality (abscess, inflammation, segmentation, cancer ...) and can provide with important indicators for the clustering of terms related to a given medical condition. Like in a previous work [25], it appears also that the anatomical localization is a secondary factor.

- *Clustering methods.* Among the two clustering methods tested, Radius and *hclust*, Radius approach appears generally to provide with better results. The main difference between the generated clusters is that *hclust* clusters are disjoint sets of terms, while Radius clusters may overlap. For this reason, the precision is better with *hclust* clusters, but the recall is then very low. With the Radius clustering, the precision is

ID	*Names of the hierarchical SMQs and of their sub-SMQs*	*Number of*		*PT*	*Evaluation*		
		levels	*s-smq*		*P*	*R*	*F*
20000060	*Cerebrovascular disorders*	3	5	198	49	70	57
	Central nervous system haemorrhages and cerebrovascular conditions			30	60	65	62
	Ischaemic cerebrovascular conditions			67	55	68	62
	Haemorrhagic cerebrovascular conditions			35	33	71	50
	Conditions associated with central nervous system			30	60	65	62
	Cerebrovascular disorders, not specified as haemorrhagic or ischaemic			18	60	37	46
20000038	*Haemorrhages*	2	2	422	36	51	42
	Haemorrhage terms (excl laboratory terms)			331	99	32	49
	Haemorrhage laboratory terms			91	60	42	50

Table 7: Evaluation figures of SMQs and their sub-SMQs.

Names of the SMQs	*Number of terms*			*Reference*			*After expertise*		
	SMQ	*clu*	*com*	*P*	*R*	*F*	*P*	*R*	*F*
Central nervous system haemorrhages and cerebrovascular conditions	23	25	15	60	65	62	84	86	85
Haemorrhage terms (excl laboratory terms)	192	95	63	99	32	49	100	33	50

Table 8: Evaluation figures of SMQs and their sub-SMQs.

less elevated but still very good, although the recall becomes interesting. Hence, the whole performances (f-measure) are better than with *hclust* clusters. This fundamental difference between the two sets of clusters leads to another observation: with the *hclust* approach, the position of ADR terms is exclusive to a given cluster, while in reality, a given ADR term may appear in different SMQs and sub-SMQs. For instance, the term *renal insufficiency* occurs in 11 SMQs and in six sub-SMQs. The Radius clustering approach, it allows the same ADR term to belong to several clusters, which suits better our applicational context.

- *Best cluster or merged clusters.* Another important difference is observed when we use the best cluster for a given SMQ (or sub-SMQ) or the merging of the n best clusters. As we already discussed, the best cluster often yields a very high precision, while the merging of the n best clusters will decrease the precision but improve the recall. In the use case where the precision is important, and this situation is the most favourable to pharmacovigilant experts, the best cluster strategy may be more convenient.

6.4 Failure analysis

We performed a qualitative and manual analysis with a pharmacovigilance expert of several clusters and sub-SMQs. We present here the analysis for two sub-SMQs: *Central nervous system haemorrhages and cerebrovascular conditions* and *Haemorrhage terms (excl laboratory terms)*.

The sub-SMQ *Central nervous system haemorrhages and cerebrovascular conditions* contains 23 aligned *PT* terms, while the associated grouping contain 25 terms. This grouping is obtained further to the merging of ten clusters. 15 terms are common to both the sub-SMQs and the grouping. This gives the following evaluation measures: P=60, R=65 and F=62. We performed the analysis of the noise (false positives) and of the silence (false negatives). Among the ten terms corresponding to false positives, the term *Locked in syndrome* is provided by five different clusters; terms *Limb traumatic amputation* and *Post traumatic osteoporosis* are provided by four clusters, which means that they have an elevated confidence. The remaining terms (*Anosognosia*, *Millard Gubler syndrome*, *Frostbite*, *Hereditary spastic paraplegia*, *Motor dysfunction* and *Spastic diplegia*) are proposed by only one cluster. A manual analysis of these false positives indicates that some of these terms (*Locked in syndrome*, *Anosognosia*, *Spastic diplegia* and *Motor dysfunction*) are related to the consequencies of stroke. In our expert's opinion, they should be included in this sub-SMQ because other consequences of stroke such as *Agnosia* or *Diplegia* are already present. The term *Hereditary spastic paraplegia* corresponds to an hereditary medical problem: therefore genetic factors are responsible for this event, they are not related to a drug. Hence, such terms should not be considered for inclusion in clusters. As a perspective, we can filter out terms meaning hereditary problems, which will improve the precision of the clusters that consist only of potential ADRs. The term *Millard Gubler syndrome* was not in the sub-SMQ in the version we exploit but it has been added later, which means that this clustering should be considered as correct. The four remaining terms (*Limb traumatic amputation*, *Frostbite*, *Chillblains* and *Post traumatic osteoporosis*) are true false positives. If we update the evaluation measures for this sub-SMQ, according to the performed analysis, it gives: P=84 and R=86. Among the eight terms corresponding to the false negatives (*Central pain syndrome*, *Paresis*, *Carotid artery aneurysm*, *Cerebral aneurysm ruptured*, *Syphilitic*, *Intracranial aneurysm*, *Carotid artery dissection*, *Amaurosis fugax* and *Dysarthria*), the semantic distance and similarity values are too large within the ontoEIM resource and for the thresholds we apply.

The sub-SMQ *Haemorrhage term (excl laboratory terms)* contains 192 aligned *PT* terms, while the associated grouping contains 65 terms. 63 terms are common to both of them.

This grouping is obtained thanks to the merging of ten clusters. The evaluation measures are: P=99, R=32 and F=49. The analysis of false positives indicates that only two terms (*Foetal maternal haemorrhage* and *Intra abdominal haemorrhage*) are in this situation. They are provided by two clusters. These two terms have been added in the newest version of the SMQs, which means that the precision for this sub-SMQ would be 100%. As for the 129 false negative terms, the semantic distance is again too large within the ontoEIM resource.

7. CONCLUSIONS

The proposed method applies the semantic distance and clustering approaches for the creation of clusters of ADR terms. Several experiments have been performed in order to test different factors which may influence the precision and recall performances. The obtained clusters provide with very good precision and we propose ways (merging of clusters) to improve the recall. The evaluation of system-generated clusters with SMQs and sub-SMQs indicates that the comparison with sub-SMQs is globally more performant because the terms are more homogeneous and semantically close within the sub-SMQs. This indicates that our approach can be helpful for the creation of fine-grained and hierarchically structured SMQs. Among the semantic distance and similarity algorithms we applied, the simplest approach (*Rada*) is the most efficient. It appears also that consideration of the term depth is not relevant. The exploitation of formal definitions, because of the currently missing information, have a negative impact. This aspect should be enriched in the future, particularly because we found out that morphology M axis provides with a very important information. Finally, the Radius approach, which generates non exclusive clusters, is more suitable for the creation of SMQs and sub-SMQs, where a given term may belong to several sets. Comparing to a previous work performed in our group, our results reach a much better level of relevance, while the manual analysis of the clusters indicates that the real results may be even better.

Future studies may lead to the identification and definition of other factors which influence the quality of clusters. For instance, the performances vary according to the SMQs and it appears that different strategies should be used for different SMQs, while currently we apply the same setting of the method to all the SMQs and sub-SMQs. More robust distances and clustering methods can also be used in the future work, as well as approaches for a better generation and evaluation of the hierarchical structures (such as hierarchical SMQs). We can also define a set of markers which would allow to filter out the irrelevant terms, such as those related to hereditary or chronic medical problems, and not induced by drugs. We can start applying the proposed method to other terminologies in order to test its portability. In this way, clusters representing the same medical condition in different terminologies (e.g., MedDRA, WHO-ART, SNOMED CT, UMLS) may also be created. Methods provided by Natural Language Processing may enrich and improve the clusters. Besides, the obtained clusters should also be evaluated through their impact on the pharmacovigilance tasks and through the exploring of the pharmacovigilance databases.

8. ACKNOWLEDGMENTS

This work was partially supported by funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) for the Innovative Medicine Initiative (IMI) under Grant Agreement nb [1150004]. The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi- protect.eu) which is a public-private partnership coordinated by the European Medicines Agency. The authors thank O. Caster, G. Declerck, R. Fescharek, R. Hill, A. Kluczka, X. Kurz, A. Jamet, MC. Jaulent, M. Lerch, N. Noren, V. Pinkston, E. Sadou, J. Souvignet and T. Vardar, but the views expressed here are those of the authors only. The authors also thank A. Périnet for her editorial assistance.

9. REFERENCES

[1] I. Alecu, C. Bousquet, and M. Jaulent. A case report: using snomed ct for grouping adverse drug reactions terms. *BMC Med Inform Decis Mak*, 8(S1):4, 2008.

[2] A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*, 54(4):315–21, 1998.

[3] O. Bodenreider. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. In *AMIA Annu Symp Proc*, pages 45–9, 2009.

[4] C. Bousquet, C. Henegar, A. Lillo-Le Louët, P. Degoulet, and M. Jaulent. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563–71, 2005.

[5] E. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–17, 1999.

[6] CIOMS. Development and rational use of standardised MedDRA queries (SMQs): Retrieving adverse drug reactions with MedDRA. Technical report, CIOMS, August 2004.

[7] R. A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett, and L. Brochu. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield, 1993.

[8] M. Dupuch, L. Trinquart, I. Colombet, M.-C. Jaulent, and N. Grabar. Exploitation of semantic similarity for adaptation of existing terminologies within biomedical area. In *Reuse and adaptation of terminologies and ontologies (WS EKAW)*, 2010.

[9] R. Fescharek, J. Kübler, U. Elsasser, M. Frank, and P. Güthlein. Medical dictionary for regulatory activities (MedDRA): Data retrieval and presentation. *Int J Pharm Med*, 18(5):259–269, 2004.

[10] M. Hauben and A. Bate. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov Today*, 14(7-8):343–57, 2009.

[11] J. Iavindrasana, C. Bousquet, P. Degoulet, and M. Jaulent. Clustering who-art terms using semantic distance and machine algorithms. In *AMIA Annu Symp Proc*, pages 369–73, 2006.

[12] M. Jaulent and I. Alecu. Evaluation of an ontological

resource for pharmacovigilance. In *Stud Health Technol Inform*, pages 522–6, 2009.

[13] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Research in Computational Linguistics*, pages 19–33, 1997.

[14] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. 1998.

[15] D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

[16] P. Lord, R. Stevens, A. Brass, and C. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[17] G. Merrill. The meddra paradox. In *AMIA Annu Symp Proc*, pages 470–4, 2008.

[18] R. Meyboom, M. Lindquist, A. Egberts, and I. Edwards. Signal selection and follow-up in pharmacovigilance. *Drug Saf*, 25(6):459–65, 2002.

[19] F. Mougin, M. Dupuch, and N. Grabar. Improving the mapping between MedDRA and SNOMED CT. In *AIME*, 2011. To appear.

[20] P. Mozzicato. Standardised meddra queries: their role in signal detection. *Drug Saf*, 30(7):617–9, 2007.

[21] P. Nadkarni and J. Darer. Determining correspondences between high-frequency MedDRA concepts and SNOMED: a case study. *BMC Med Inform Decis Mak*, 10:66, 2010.

[22] H. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *IEEE Eng Med Biol Proc*, pages 623–8, 2006.

[23] NLM. *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland, 2008. www.nlm.nih.gov/research/umls/.

[24] R. Pearson, M. Hauben, D. Goldsmith, A. Gould, D. Madigan, D. O'Hara, S. Reisinger, and A. Hochberg. Influence of the meddra hierarchy on pharmacovigilance data mining results. *Int J Med Inform*, 78(12):97–103, 2009.

[25] D. Petiot, A. Burgun, and P. Le Beux. Modelisation of a criterion of proximity: Application to medical thesauri. In *Medical Informatics Europe*, pages 149–52, 1996.

[26] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on systems, man and cybernetics*, 19(1):17–30, 1989.

[27] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence*, 2004.

[28] K. Spackman and K. Campbell. Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. In *Journal of American Medical Informatics Association (JAMIA)*, pages 740–744, 1998.

[29] M. Stearns, C. Price, K. Spackman, and A. Wang. Snomed clinical terms: overview of the development process and project status. In *AMIA*, pages 662–666, 2001.

[30] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM*, 1993.

[31] M. Warin, H. Oxhammar, and M. Volk. Enriching an ontology with wordnet based on similarity measures. In *MEANING-2005 Workshop*, 2005.

[32] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of Associations for Computational Linguistics*, pages 133–138, 1994.

[33] J. Zhong, H. Zhu, J. Li, and Y. Yu. Conceptual graph matching for semantic search. In *10th International Conference on Conceptual Structures, ICCS2002, LNCS 2393, Springer Verlag*, pages 92–106, 2002.