Semantic distance-based creation of clusters of pharmacovigilance terms and their evaluation

Marie Dupuch (1,2,3)[1] and Natalia Grabar (1)

(1) CNRS UMR 8163 STL; Université Lille 1&3, F-59653 Villeneuve d'Ascq, France
(2) INSERM, U872, Paris, F-75006, France
(3) Viseo-Objet Direct, 4, avenue Doyen Louis Weil, F-38000 Grenoble, France

Abstract

Background: 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 using statistical algorithms and groupings of *ADR* terms from the MedDRA (Medical Dictionary for Drug Regulatory Activities) terminology. Standardized MedDRA Queries (*SMQs*) are the groupings which become a standard for assisting the retrieval and evaluation of MedDRA-coded *ADR* reports worldwide. Currently 84 *SMQs* have been created, while several important safety topics are not yet covered. Creation of *SMQs* is a long and tedious process performed by the experts. It relies on manual analysis of MedDRA in order to find out all the relevant terms to be included in a *SMQ*. Our objective is to propose an automatic method for assisting the creation of *SMQs* using the clustering of terms which are semantically similar.

Methods: The experimental method relies on a specific semantic resource,

[1]Corresponding author: dupuchm@hotmail.fr; CNRS UMR 8163 STL; Université Lille 1&3, F-59653 Villeneuve d'Ascq, France; tel: +33 3 20 41 68 51, fax: +33 3 20 41 67 14

and also on the semantic distance algorithms and clustering approaches. We perform several experiments in order to define the optimal parameters.

Results: Our results show that the proposed method can assist the creation of *SMQs* and make this process faster and systematic. The average performace of the method is precision 59% and recall 26%. The correlation of the results obtained is 0.72 against the medical doctors judgments and 0.78 against the medical coders judgments.

Conclusions: These results and additional evaluation indicate that the generated clusters can be efficiently used for the detection of pharmacovigilance signals, as they provide better signal detection than the existing SMQs.
Keywords: Pharmacovigilance, Terminology, Clustering, Semantic Distance and Similarity, Evaluation, Signal Detection, MedDRA, SMQs

1. Introduction

During new drug development, clinical trials are performed in order to test them, to study the reaction of human subjects to them and to detect the most common adverse drug reactions (*ADRs*) and risks. However, the clinical trials involve several thousand patients at most. As a result, less common *ADRs*, although they may be severe, remain often undiscovered at the end of the clinical trials and when a drug is put on the market. Continuous surveillance of the safety topics (*i.e.*, *Haemorrhages*, *Anaphylactic shock*, *Rhabdomyolysis*, *Acute renal failure*, *Cardiac failure*) and of the use of the drugs is then necessary. It is done through pharmacovigilance activity accomplished at regional, national and international levels. This activity relies on collection and analysis of spontaneous reports submitted by health

professionals and, in some countries, by patients. Although the collection of spontaneous reports is not exhaustive [1, 2], the resulting pharmacovigilance databases are very large. To facilitate pharmacovigilance data recording and analysis, the *ADRs* from the spontaneous reports are coded using a controlled vocabulary, usually MedDRA (Medical Dictionary for Drug Regulatory Activities) [3]. Then, pharmacovigilance experts perform a manual review of these reports. More recently, in some countries, statistical data mining techniques are also applied [4, 5]. However, it was observed that because pharmacovigilance terminologies are often fine-grained (*i.e.*, MedDRA contains over 80,000 terms), the combination of multiple terms denoting similar notions (*e.g.*, *Hepatitis infectious*, *Hepatitis infectious mononucleosis*, *Hepatitis viral*) is necessary during the signal detection[2] [6, 7]. In this context, the groupings of semantically close *ADR* terms can be useful.

2. Research questions

Our objective is to propose new and efficient methods for assisting signal detection and for grouping pharmacovigilance terms. This is a poorly investigated area. More precisely, we propose to rely on semantic distance and clustering methods, which we assume to be likely to produce relevant clusters because semantically close terms may be detected and grouped together with these methods. We chose the MedDRA terminology because it is used worldwide in the pharmacovigilance domain.

In the remainder of this article, we first present the related work. We

[2]Pharmacovigilance signal is a new or unknown relation between a drug and an *ADR*.

Level	*Expanded form*	*Terms examples*	*Nb Terms*
SOC	System Organ Class	*Cardiac disorders*	26
HLGT	High Level Group Terms	*Cardiac arrhythmias*	332
HLT	High Level Terms	*Rate and rhythm disorders*	1,688
PT	Preferred Terms	*Bradycardia*	18,209
LLT	Lowest Level Terms	*Bradycardia*, *Bradycardiac tendency*, *Reflex bradycardia*...	66,587
Total			86,842

Table 1: Five hierarchical levels of MedDRA: terms examples and number per level.

then describe material and methods we propose for testing and evaluating our approach. In order to better assess the proposed method relevance, special attention is paid to the evaluation of the generated clusters. We finally discuss the obtained results and conclude with some perspectives.

3. Related work

3.1. Grouping pharmacovigilance terms

The MedDRA terms are structured into five hierarchical levels (Table 1): System Organ Class (*SOC*), High Level Group Term (*HLGT*), High Level Term (*HLT*), Preferred Term (*PT*) and Low Level Term (*LLT*). The highest level *SOC* is related to human body organs (such as *Cardiac disorders*, *Immune system disorders*, *Eye disorders* or *Psychiatric disorders*), while other levels provide hierarchical subsumption of terms from the corresponding lower level. For instance, the *PT Brachycardia* term is subsumed by its *HLT* term *Rate and rhythm disorders*. The *LLT* terms have a special

place [8]: they can be synonyms of their *PTs* or they can convey more specific notions (*Bradycardiac tendency* or *Reflex bradycardia* in Table 1).

In the existing studies, grouping of pharmacovigilance terms is based either on the MedDRA terminology structure or on the use of derived resources. The first type of approach for term grouping is based on the hierarchical structure of MedDRA, that is the *HTL, HLGT* or *SOC* levels [9, 10]. It considers together terms which have common hierarchical parents or ancestors and which also share some common semantic features. However, it was observed that some safety topics are transverse to these hierarchical levels of MedDRA, which means that relevant terms can belong to different *HLTs, HLGTs* or *SOCs*. This fact led to the development of the Standardized MedDRA Queries (*SMQs*) containing MedDRA terms in connection with a safety topic [11] and independently from the *SOCs* of these terms. For example, the *Haemorrhages SMQ* contains an aggregation of the MedDRA terms related to bleeding in all parts of the body, and thus in a broad set of *SOCs* (*Vascular disorders*, *Gastrointestinal disorders*, *Reproductive system and breast disorders*...). The *SMQs* are developed by international groups of experts looking manually through the MedDRA terminology in order to detect relevant terms to each *SMQ*.

A specific resource, called ontoEIM[3] [12], has been created by projecting MedDRA and WHO-ART (WHO Adverse Reaction Terminology) terminologies on the SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) terminology [13]. This projection was performed using the

[3]ontoEIM stands for *ontologie des Événements Iatrogènes Médicamenteux* (*ontology of drug-induced events*)

UMLS (Unified Medical Language System) [14], which already merges and partially aligns these terminologies. ontoEIM was then used to perform hierarchical subsumption of terms and to group them together [12, 15]. Precision observed was high while recall was extremely low, which may be explained by the fact that hierarchical subsumption seems to be irrelevant for the creation of groupings of pharmacovigilance terms. In other experiments, the ontoEIM resource has been exploited with a semantic distance approach and applied to a subset of MedDRA [16] and WHO-ART terms [17]. In the WHO-ART related experiment, the obtained groupings demonstrated interesting results, because several semantic relationships were indeed detected (synonyms, antonyms, physiological functions or abnormalities, associated symptoms, abnormal laboratory tests, pathologies and their causes, close anatomical localizations, degrees of severity, and heterogeneous groupings), although these groupings have not been compared with the *SMQs*. Therefore, we propose to further adapt and evaluate semantic distance measures for this task.

3.2. Semantic distance and similarity

Semantic distance and similarity measures indicate the semantic relatedness between two words or expressions. In the following, we call them *semantic distance* measures. The advantage of these measures is that they quantify semantic relatedness and provide numerical values, which can feed other computational applications. Several approaches exist to compute them. Typically, these measures are distinguished according to whether they rely on corpora or on tree-structured resources (lexical networks, terminologies, ontologies...) and/or whether they are path-based or node-based. In Table

Measures	*Resource*	*SP*	*NCP*	*Depth*	*Density*	*IC*
Rada [18]	MeSH	+	-	-	-	-
Sussna [19]	WordNet	+	-	+	+	-
Zhong [20]	Conceptual graphs	+	+	+	-	-
Wu & Palmer [21]	WordNet	+	+	+	-	-
Jarmasz & Szpakowicz [22]	Roget's Thesaurus	+	-	-	-	-
Resnik [23]	WordNet	-	+	+	-	+
Leacock & Chodorow [24]	WordNet	+	-	-	-	-
Jiang & Conrath [25]	WordNet	+	+	+	+	+
Lin [26]	WordNet	-	+	+	-	+
Hirst & St Onge [27]	WordNet	+	-	-	-	-
Steichen et al. [28]	Medical ontology	-	+	+	+	+
Cho [29]	WordNet	+	+	+	+	+
Yang [30]	WordNet	-	-	+	-	-

Table 2: Most frequently used semantic similarity and distance algorithms. *SP* stands for the shortest path, *NCP* stands for the nearest common parent, *IC* stands for information content. The + means that the technique mentioned in a given column is used in a given reference from the first column.

2, we indicate the most frequently used semantic distance measures. Measures from the first series [18, 19, 20, 21, 22] are path-based. The first and the simplest measure of the kind was proposed by Rada [18]: it relies on tree-structured resources and counts the edges between two entities. The measures from this set use only hierarchical *is-a* relations. As indicated in Table 2, path-based approaches may take into account other factors such as depth, nearest common parent or density.

The second set of measures [23, 24, 25, 26] are node-based. They rely on corpora, used with [24] or without tree-structured resources. Semantic information content, which allows semantic relatedness to be computed between two nodes (terms or expressions), is then associated with the nodes. It can rely on features such as frequency observed in corpora, semantic specificity and depth in a tree-structured resource.

The common feature of the third series of measures [27, 28, 29, 30] is that they use not only hierarchical relations, but also other types of relations (such as *treatment of*, *causes*, *finding site of*, *associated morphology of*, etc.). Such relations are indeed available in some terminologies and ontologies, such as SNOMED CT [13], FMA (Foundational Model of Anatomy) [31] or WordNet [32]. For instance, in the SNOMED CT, the terms *renal insufficiency* and *kidney* belong respectively to *Disorders* and *Body structure* hierarchies and are connected by the *finding site of* relation: *renal insufficiency* is localized in *kidney*. Because the meaning of non-hierarchical relations may be very different, these relations have to be ranked and some paths (*i.e.*, from non-hierarchical to hierarchical relations) may be forbidden.

Finally, it is important to note that the existing similarity measures, even

if they have been designed in other contexts, may be adapted to biomedical data and terminologies, *i.e.*, MeSH (Medical Subject Headings), SNOMED CT, GO (Gene Ontology) [33, 34, 35, 36].

3.3. Term clustering

The objective of clustering methods is to organize similar objects (*i.e.*, terms) within homogeneous groups, while dissimilar objects (or terms) will belong to different groups. We distinguish three categories of clustering algorithms, according to the types of generated clusters: (1) disjoint clusters (a given object may belong to one cluster only), (2) non disjoint clusters (a given object may belong to more than one cluster), and (3) hierarchical clusters considered as non disjoint when viewed through the dendrogram (smaller clusters are included into the larger clusters) or disjoint once the dendrogram is cut. We describe some of the algorithms in the following.

Disjoint clustering is performed with algorithms such as *k-means* [37], *k-medoids* and *PAM* [38]. They are adapted to large data processing. With these algorithms, it is necessary to indicate the number of clusters to be generated. Their specificity (number of clusters to be generated must be indicated and disjoint character of the clusters) is not suitable for our present study. Non disjoint clustering is performed with so-called, fuzzy or soft algorithms. Fuzzy algorithms (*fuzzy c-means* [39], *fuzzy c-medoids* [40] or *axial k-means* [41]) state the degree up to which an object belongs to each concerned cluster. The difficulty with these algorithms is that they require to set up thresholds, which may be a difficult step. The few existing soft clustering algorithms (*i.e.*, PoBOC [42], OKM [43] and *Radius* [44]) also allow an intersection between generated clusters but without specifying the

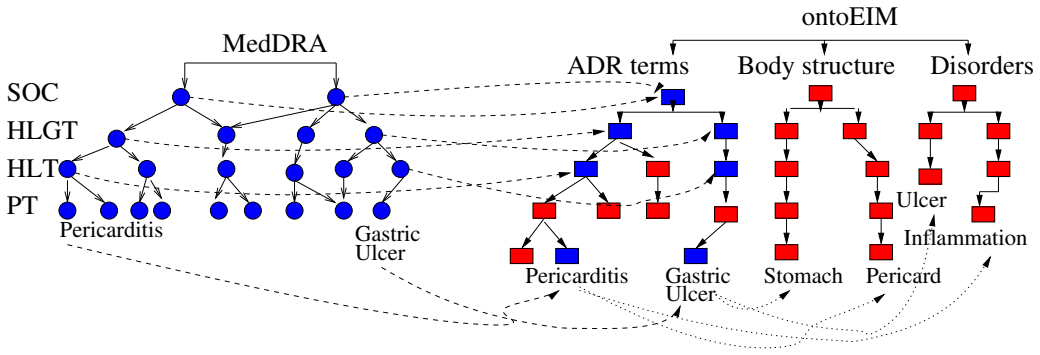


Figure 1: Projection of MedDRA on the SNOMED CT terminology results in the creation of the ontoEIM semantic resource.

relevance degree of each entity to a given cluster. Finally, several hierarchical clustering algorithms have been proposed (AGNES [45, 46], BIRCH [47], CURE [48] and DIANA [46]). With these algorithms, it is not necessary to set up the number of classes, which makes them easy to apply. We assume non disjoint soft clustering and hierarchical clustering may be suitable for our purpose, and therefore propose to apply and test them in our study.

4. Material

Our material is specific to the pharmacovigilance area and consists of terms from the MedDRA terminology [3], the reference groupings of terms, and a pharmacovigilance database.

ontoEIM resource

The MedDRA-derived ontoEIM resource [12] has been created using the UMLS (version 2010AB) in which some MedDRA (46%) and SNOMED CT terms are already aligned. This resource can be easily updated with new releases of the UMLS. ontoEIM improves MedDRA *PT* terms representation.

The first advantage is that the MedDRA term structure becomes parallel to their structuring in SNOMED CT, which makes it more fine-grained: a SNOMED CT-derived hierarchy of the MedDRA terms contains terms and intermediate hierarchical levels absent in MedDRA. Thus, on the right graph of Figure 1, MedDRA terms are blue nodes while all other terms (red nodes) are provided only by SNOMED CT. Another advantage is that *ADR* terms may be decomposed into semantic primitives. In our study, we decompose them into two primitives (*Disorders* and *Body structure*) from the SNOMED CT, as exemplified on Figures 1 and 3: for instance, *Gastric ulcer* is decomposed into *Ulcer* and *Stomach*.

ontoEIM (MedDRA *PT* terms, their structure and semantic decomposition) is our main material for creating MedDRA *ADR* term groupings. We use *PT* terms because they are used for the coding of pharmacovigilance reports and also form the core part of the *SMQs*, which may be further extended with their *LLT* terms.

Standardized MedDRA Queries (SMQs)

We use 84 existing *SMQs* (2011 version), which cover several safety topics such as *Haemorrhages*, *Anaphylactic shock*, *Rhabdomyolysis*, *Acute renal failure*, *Cardiac failure*. *SMQs* contain terms which are distinguished according to whether they belong to the narrow or broad version of the *SMQs*. On the *Acute renal failure SMQ* example, the narrow version contains main terms which are strongly associated to this *ADR* (*i.e.*, *Renal failure, Dialysis, Renal impairment, Haemodialysis*), while the broad version includes also secondary terms (*i.e.*, *Urine output decreased, Nephritis, Renal transplant, Renal tubular disorder*) which become meaningful when they are combined

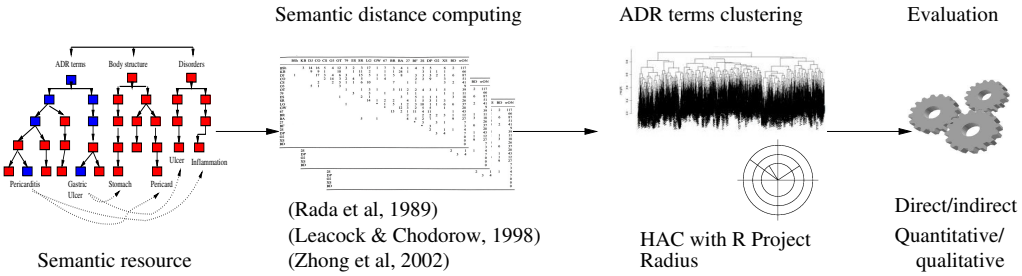


Figure 2: Main steps of the method for the grouping of the *ADR* terms with semantic distance algorithms.

between them or with main terms. *SMQs* are used as the gold standard to evaluate generated *ADR* term groupings.

FDA AERS database

The FDA AERS[4] is the official database of the *ADRs* spontaneous reports in the United States. It is publicly available. AERS contains over 2 million reports coded with the MedDRA *PT* terms. We use this database when evaluating the generated groupings within the signal detection context.

5. Methods

The proposed method is organized into three main steps (Figure 2): (1) computing the semantic distance between the *ADR* terms, (2) clustering the *ADR* terms, and (3) evaluation of the obtained clusters. Implementation is done in Perl and *R*[5] languages. In previous work [49, 44], we started to

[4] *www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm*: Adverse Event Reporting System of the Food and Drug Administration

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

exploit such methods, while in the current experiments we further adapt them to the creation of groupings of the *ADR* terms and perform a detailed evaluation within direct (theoretical) and indirect (applicational) contexts.

Computing the semantic distance between ADR PT terms

Semantic distance is computed between the 7,629 *PT* MedDRA terms present in the ontoEIM resource. We use only the *PT* terms because they constitute the *SMQs* and they are used for the coding of pharmacovigilance case reports worldwide. The computing of semantic distance is performed for every possible pair of terms to build the symmetric matrices 7,629*7,629. During this step, we apply three measures to compute the semantic distance between two *ADR* terms *t1* and *t2*. These measures have been chosen because they are suitable for tree-structured resources, like ontoEIM, and they involve different factors (the shortest path, the maximal depth and the nearest common parent):

- the *Rada* approach [18] computes the distance and relies on the shortest path *sp* detection, which corresponds to the sum of this shortest path edges: $sp(t1, t2)$

- the *LCH* Leacock and Chodorow's approach [24] computes the similarity and relies on the shortest path *sp* and on the maximal depth *MAX* found within the resource (*MAX*=14 within the ontoEIM): $-log[\frac{sp(t1,t2)}{2*MAX}]$

- the *Zhong* approach [20] computes distance and relies on absolute depth *depth* of terms and on their nearest common parent *ncp*. According to [20], the nearest common parent is the hierarchical parent node

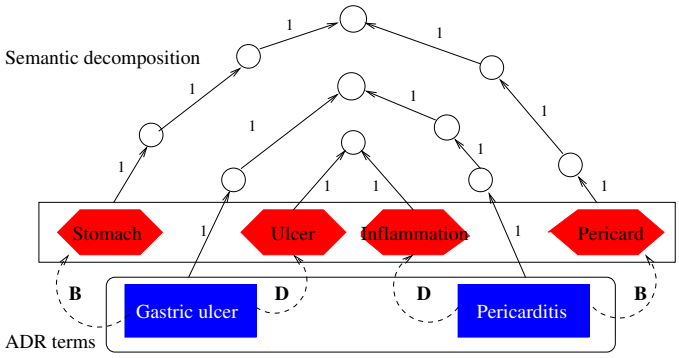


Figure 3: Computing the shortest path *sp* between two ADR terms (*Pericarditis* and *Gastric ulcer*) directly and through their semantic decomposition.

which is the closest to two terms $t1$ and $t2$. The milestone value m is computed first for each term: $m(t) = \frac{1}{k^{depth(t)+1}}$, where t is a term, *depth* its absolute depth within a terminology and $k = 2$ (normalization coefficient). Then, the distance between two terms is computed as: $2 * m(ncp(t1, t2)) - (m(t1) + m(t2))$.

Semantic distance is computed between the *ADR PT* terms but also between their semantic primitives provided by the D (Disorders) and B (Body structure) axes. Semantic decomposition is exploited to make the *ADR* term representation fine-grained [50]. Figure 3 illustrates how the shortest paths *sp* are computed between two *ADR* terms (*Gastric ulcer* and *Pericarditis*) and between their semantic primitives (axes D and B). The edge weight is set to 1 because all relations are of the same kind (hierarchical), and each shortest path value corresponds to the sum of its edge weights. For this pair of terms, we obtain the following values: $sp_{ADR} = 5$, $sp_B = 6$ and $sp_D = 2$.

The semantic distance computing is then performed according to the three measures described above: *Rada*, *LCH* and *Zhong*. These seman-

tic distances sd are then applied to compute the unique distance between the ADR terms: $\forall_k \in \{Rada, LCH, Zhong\} \frac{\sum_{i \in \{ADR,D,B\}} W_i * sd_k(t1_i, t2_i)}{\sum_{j \in \{ADR,D,B\}} W_j}$, in which $\{Rada, LCH, Zhong\}$ are the tree semantic measures, $\{ADR, D, B\}$ respectively correspond to terms meaning the ADR, axis Disorders D and axis Body structure B; $t1$ and $t2$ are two ADR terms; W is the coefficient associated with each of the three terms; and sd is the semantic distance computed on a given axis and with a given semantic measure. Several experiments are performed:

1. Semantic decomposition: (1) the semantic decomposition is taken into account and the semantic distance is computed on three axes (ADR, B, D), or (2) the semantic decomposition is not taken into account and the semantic distance is computed on the axis of $ADRs$ only;
2. Coefficient W put on the ADR terms axis and on D and B axes are set either to 1 or to 2 and all the possible combinations are tested to assess the semantic decomposition impact.

Further to the application of this method, symmetric matrices 7,629*7,629 are built. They contain semantic distances between $ADR\ PT$ terms.

Clustering the ADR PT terms

Once the distances are computed, we use them to generate clusters of terms. Because a given ADR term may appear in different $SMQs$ (*i.e.*, *renal insufficiency* occurs in 11 $SMQs$), we have to generate non disjoint clusters. Among the clustering methods presented in section 3.3, we apply hierarchical classification HAC and non disjoint clustering with R Radius approach:

- the *HAC* hierarchical ascendant classification is performed using the *R Project* tools[6]. This method first chooses the best centers for clusters and then builds the term hierarchy by progressively merging smaller clusters to obtain only one big cluster. The final dendrogram is segmented into n clusters, in which n is tested within the interval [100; 200; 300; ... 7,000].

- the *R* radius approach, with which every *ADR* term is considered as a possible center of a cluster and its closest terms within a given distance are clustered together with it. We test several semantic distance values within the following intervals: two singletons 2 and 3 for *Rada*, [0; 5.059] for *LCH* and [0; 0.49] for *Zhong* (the last two upper values are the maximal values obtained within the ontoEIM resource). Moreover, when terms of a small cluster are included in a bigger cluster, the small cluster is removed. In addition, when two clusters have at least 80% of common terms they are merged.

Evaluation

We perform several kinds of evaluation, among which we distinguish direct and indirect evaluation, and also quantitative and qualitative evaluation. Quantitative evaluation is done against the gold standard and a baseline. It is usually measured with precision and recall values, while qualitative evaluation requires experts' opinion. As for direct (or theoretical) evaluation, it assesses the correctness of term pairs provided by the semantic distance measures through their comparison with the similar data manually created

[6] *http://www.r-project.org*

by the experts, while indirect (or application-based) evaluation considers the same results, but through their relevance to the aimed applications. In our experiments, the applications are related to the creation of *SMQs* and to the signal detection.

Comparison of rating of term pairs with human judgment. This evaluation objective is to analyze whether the measures (and the resource) can reproduce the expert opinion on the semantic relatedness between terms. In this evaluation, we rely on the reference data provided by a previous study [51]. These data contain 30 term pairs manually annotated by three medical doctors and nine medical coders. The annotators were asked to rate the term pairs on a scale [1, 2, 3, 4], where 1 stands for semantically different terms and 4 for semantically identical terms. The correlation among the annotators is 0.68 for medical doctors and 0.78 for coders. Some of these 30 pairs could not be used in our study because:

- eleven terms from these pairs are not MedDRA terms (*i.e.*, *myocardium*, *calcification*, *lymphoid hyperplasia*),
- seven other terms do belong to MedDRA but are not aligned with SNOMED CT (*i.e.*, *hyperlipidemia*, *cholangiocarcinoma*, *infarctus* or *pulmonary fibrosis*).

These two constraints reduce the number of term pairs to 14 (the first two columns in Table 4). Evaluation against the expert-rated pairs of terms is done following four steps:

1. We consider the computed similarity scores with each applied similarity measure;

2. We rate the term pairs according to their similarity scores;
3. In order to reduce the computed similarity scores to the manually applied scale [1, 2, 3, 4] and to make this evaluation feasible, we apply the non-parametric Spearman rank correlation coefficient;
4. We compute the correlation between human and automatically computed scores. It is considered that the correlation 0-0.5 is very low, 0.5-0.7 low, 0.7-0.8 moderate, 0.8-0.9 high and 0.9-1 very high.

Comparison of clusters with the reference data (84 SMQs). Quantitative evaluation of the generated clusters is performed by comparing them with the 84 *SMQs*. Three classical measures are computed: precision P (number of relevant clustered terms divided by the total number of clustered terms), recall R (number of relevant clustered terms divided by the number of terms in the corresponding *SMQ*) and F-measure F (P and R harmonic mean). The association between the *SMQs* and the clusters relies on F-measure values. We evaluate the generated clusters against narrow (main *ADR* terms) and broad (all the terms) versions of the *SMQs*.

Comparison of clusters with the baseline (46 SMQs). For the baseline, we chose the most frequently used approach for MedDRA term grouping, which relies on the MedDRA hierarchical structure, that is hierarchical subsumption of *PTs* through the *HLT* MedDRA level [9, 10, 52]. Among the 1,688 *HLTs* and 84 *SMQs*, 46 of them have direct (*Thrombocytopenias (SMQ)* and *Thrombocytopenia (HLT)*) or non ambiguous correspondences (*Renal failure and impairment (SMQ)* and *Acute renal failure (HLT)*). We use these 46 *SMQs* as our baseline reference. These 46 *SMQs* are a subset of the whole set

	Drug j	Other drugs	
ADR_i	n_{ij}	$n_{i\bar{j}}$	n_i
Other $ADRs$	$n_{\bar{i}j}$	$n_{\bar{i}\bar{j}}$	$n_{\bar{i}}$
	n_j	$n_{\bar{j}}$	n

Table 3: Statistical test for the signal detection in a pharmacovigilance database: n is the total number of $\{drug, ADR\}$ pairs, n_{ij} is the number of reports involving ADR_i and Drug_j, n_i is the marginal count involving ADR_i, n_j is the marginal count involving and Drug_j

of 84 *SMQs*. The evaluation measures are, as previously described: precision P, recall R and F-measure F.

Analysis of clusters with an expert. Qualitative evaluation of clusters is performed with a medical expert. The expert is asked to provide a judgment on the content of clusters and its relevance to a given safety topic. The objective of this evaluation is to propose failure analysis through the study of false positive and false negative terms.

Evaluation through signal detection. One last evaluation is performed with the freely available FDA AERS database in order to assess the suitability of the generated clusters for the signal detection. Several statistical tests exist[7], *i.e.*, EBGM applied by the FDA, IC by the World Health Organiza-

[7]EBGM (Empirical Bayes Geometric Mean), IC (information component), PRR (proportional reporting ratio) and ROR (reporting odds ratio) are mathematical tools for identifying signals of disproportional reporting of suspected *ADR*s in association with particular drugs.

tion, PRR by the UK Medicines Control Agency, ROR by the Netherlands authorities, etc. With these tests, a signal appears when the number of observed cases is higher than the number of expected cases. The threshold can be established with a mathematical model, such as in Table 3, in which the four variables (n_{ij}, $n_{\bar{i}j}$, $n_{i\bar{j}}$ and $n_{\bar{i}\bar{j}}$) imply all the drugs and all the *ADRs* in a pharmacovigilance database. We apply the EBGM test, used with the FDA database, to evaluate clusters generated with our methods. As an example, if EBGM = 3.9 (*i.e.*, with the pair {*acetaminophen*, *hepatic failure*}), this means that this {*drug*, *ADR*} pair occurs in the database 3.9 times more frequently than expected. We compute a 90% confidence interval. We use the EB05 criterion which had to be superior or equal to a threshold value of 2. The data-mining signal EB05>= 2 means that the pair {*drug*, *ADR*} occurred at least twice as often as expected. Such threshold guarantees that potential signals are likely to be correct. For this evaluation, we randomly select 19 active ingredients (*acetylsalicylic acid SRT, Eloxatin, Fentanyl citrate, Flovent, Humulin N, isosorbide mononitrate, Januvia, Leflunomide, Lisinopril, Lorazepam, Methadone HCL, Methotrexate sodium, Nevirapine, Pravachol, Soliris, Sutent, Torsemide, Vioxx, Zolpidem*) and we compute the EB05 values for the generated clusters and the corresponding *SMQs* or *HLTs*. Then we analyze the variability values obtained between the clusters and *SMQs/HLTs* by computing the regression line by the method of least squares (linear equation $y = ax + b$) and the coefficient of determination R^2. The coefficient of determination R^2 may vary between 0 (no correlation) and 1 (perfect correlation).

6. Results

The 7,629 *ADR PT* terms from ontoEIM have been processed with the three semantic measures and the two clustering algorithms outlined in previous section. For semantic measures, the best thresholds are: 2 for *Rada*, 4.10 for *LCH* and 0.20 for *Zhong*. With higher thresholds, the generated clusters are too large and become meaningless. The *Rada* measure outperforms the other two measures. With the *HAC* algorithm, the best results are obtained with 300 classes. In addition, we obtain better results when semantic decomposition is not applied (*ADR* terms only). The results we present and discuss in the following are obtained with our optimal parameters: the *Rada* semantic measure, no semantic decomposition, and the *Radius* clustering algorithm with a threshold set to 2. We then obtain 2,931 clusters. The number of terms per cluster varies between 2 and 546 (mean=17). The evaluation against term pairs manually rated by several experts indicates our methods provide results very close to human judgment. The evaluation with a signal detection protocol, although we did not generate the exact content of the *SMQs*, indicates that the clusters seem to be suitable for the signal detection task. The manual evaluation of the clusters by a medical expert shows that relevant terms may be missing in the *SMQs* [9, 10], some of which can be found with our methods. We detail these results in the following section.

Comparison of rating of term pairs with human judgment

Among all the generated pairs with the 7,629 MedDRA terms, we evaluated 14 term pairs through the comparison of the computed similarity scores with those provided by human experts in a previous study [51]. Scores ob-

Medical Term Pairs		ontoEIM					
		Rada		LCH		Zhong	
Term 1	Term 2	score	rank	score	rank	score	rank
Renal failure	Kidney failure	1	1	3.33	1	0	1
Abortion	Miscarriage	2	2	2.639	2	0.013	3
Brain tumor	Intracranial haemorrhage	4	3	1.94	3	0.001	2
Pulmonary embolus	Myocardial infarction	5	4	1.72	4	0.046	5
Multiple sclerosis	Psychosis	6	5	1.54	5	0.105	9
Diarrhea	Stomach cramps	7	6	1.38	6	0.120	10
Congestive heart failure	Pulmonary edema	7	6	1.38	6	0.052	6
Carpal tunnel syndrome	Osteoarthritis	7	6	1.38	6	0.021	4
Mitral stenosis	Atrial fibrillation	7	6	1.38	6	0.013	3
Metastasis	Adenocarcinoma	8	7	1.25	7	0.056	7
Appendicitis	Osteoporosis	9	8	1.13	8	0.057	8
Peptic ulcer disease	Myopia	11	9	0.93	9	0.241	12
Depression	Cellulitis	11	9	0.93	9	0.233	11
Schizophrenia	Delusion	13	10	0.76	10	0.241	13

Table 4: Evaluation of the obtained scores for the 14 term pairs by comparing them with the scores described in (Pedersen et al., 2007).

Measure	Medical doctors	Coders
Rada	0.72	0.78
LCH	0.72	0.78
Zhong	0.46	0.59

Table 5: Correlation results for medical doctors and coders.

tained with the three similarity measures (*Rada, LCH* and *Zhong*) for each term pair are provided in Table 4. Term pairs are sorted according to their *Rada* similarity scores: at the top of the table, term pairs have a strong semantic association (*e.g.*, {*Renal failure*, *Kidney failure*}, {*Abortion*, *Miscarriage*}), while at the bottom of the table, terms are judged to be semantically dissimilar (*e.g.*, {*Depression*, *Cellulitis*}, {*Peptic ulcer disease*, *Myopia*}). In Table 5, we indicate the correlations between the scores obtained with our methods and those provided by the experts in [51]. According to the grid mentioned in the Methods section, the correlation we obtain with *LCH* and *Rada* measures is moderate and close to high: 0.72 against medical doctors judgments and 0.78 against medical coders judgments. With the *Zhong* measure, this correlation is very low (0.46) and low (0.59). On the whole, we obtain better correlations than those reported in previous study: 0.35 and 0.50, respectively [51]. This means that the similarity scores computed with the ontoEIM resource and the applied method are quite close to human judgment, especially to the judgment of medical coders. We assume, this result may also have a positive impact on other evaluations.

Comparison of the generated clusters with the reference data (84 SMQs)

The generated clusters have been evaluated against the 84 *SMQs*. As indicated on Figure 4, the applied method provides a good precision for several *SMQs*, but the recall remains low because the generated clusters are smaller than the corresponding *SMQs*: they usually correspond to different facets of the *SMQs*. The average performance is: P=52, R=25, F=31, although there is great variability between the clusters. Some of the clusters are distinguished by their high precision (*i.e.*, *Haemorrhages, Cardiac arrhythmias,*

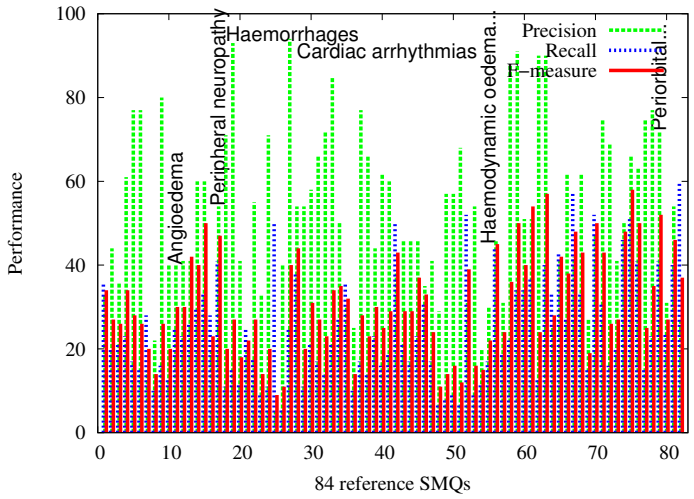


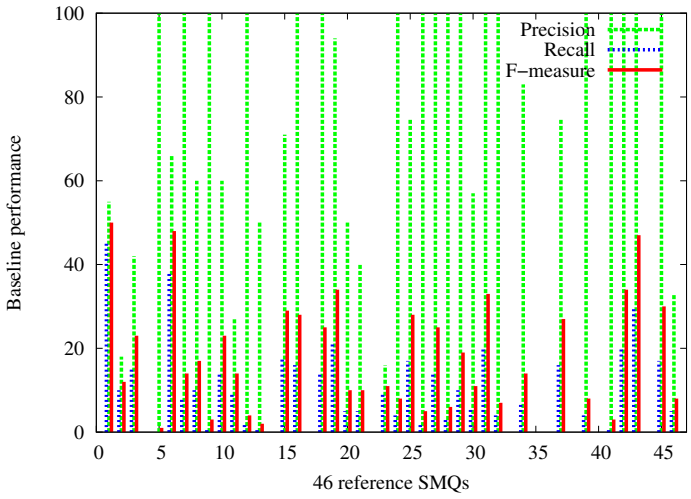
Figure 4: Precision, recall and F-measure values of the generated clusters through their comparison with the SMQs.

Hypertension) or high recall (*i.e.*, *Periorbital and eyelid disorders, Taste and smell disorders*), others by their low precision (*i.e.*, *Taste and smell disorders, Hyponatraemia*) or low recall (*i.e.*, *Anaphylactic reaction, Agranulocytosis*).

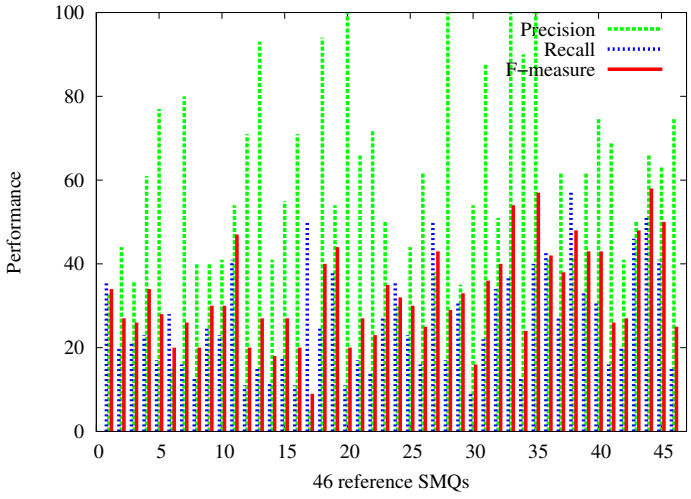
In Table 6, we present in detail the content of the cluster which corresponds to the *SMQ Anaphylactic shock*. In the first column, we indicate the MedDRA terms, in the second we specify whether these terms belong or not to the *SMQ* (if they do, we indicate the version, broad or narrow, of the *SMQ*), we then mention whether the terms are aligned with the corresponding SNOMED CT terms and whether they are included in the generated cluster. The *SMQ* and the cluster have 7 common terms, while 15 more terms from the cluster are not included in the *SMQ*. We further analyze these data below.

MedDRA term	*SMQ*	Aligned	Cluster
Anaphylactic reaction	Narrow	-	-
Anaphylactic shock	Narrow	-	-
Anaphylactic transfusion reaction	Narrow	+	+
Anaphylactoid reaction	Narrow	+	-
Anaphylactoid shock	Narrow	-	-
Circulatory collapse	Narrow	-	-
Shock	Narrow	+	-
Acute prerenal failure	Broad	-	-
Acute respiratory failure	Broad	+	+
Anuria	Broad	+	+
Blood pressure immeasurable	Broad	-	-
Cerebral hypoperfusion	Broad	-	-
Grey syndrome neonatal	Broad	+	-
Hepatic congestion	Broad	+	-
Hepatojugular reflux	Broad	+	-
Hepatorenal failure	Broad	-	-
Hypoperfusion	Broad	+	-
Jugular vein distension	Broad	-	-
Multi-organ failure	Broad	-	-
Myocardial depression	Broad	-	-
Neonatal anuria	Broad	-	-
Neonatal multi-organ failure	Broad	-	-
Neonatal respiratory failure	Broad	+	+
Organ failure	Broad	-	-
Propofol infusion syndrome	Broad	-	-
Renal failure	Broad	+	+
Renal failure acute	Broad	+	+
Renal failure neonatal	Broad	-	-
Respiratory failure	Broad	+	+
Acute pulmonary oedema	-	+	+
Allergic transfusion reaction	-	+	+
Cardio respiratory arrest	-	+	+
Cardio respiratory arrest neonatal	-	+	+
Chronic respiratory failure	-	+	+
Crush syndrome	-	+	+
Haemolytic transfusion reaction	-	+	+
Haemolytic uraemic syndrome	-	+	+
Hepatorenal syndrome	-	+	+
Mountain sickness acute	-	+	+
Neonatal respiratory arrest	-	+	+
Polyuria	-	+	+
Pulmonary renal syndrome	-	+	+
Respiratory arrest	-	+	+
Transient tachypnoea of the newborn	-	+	+
Total	29	28	22

Table 6: Content of the cluster which corresponds to the SMQ *Anaphylactic shock*.



(a) Baseline results



(b) Results generated by the proposed method

Figure 5: Comparison of the results generated by the proposed method with the baseline results.

SMQs	*Terms number*			*Reference*			*Expert*		
	SMQ	*clu*	*com*	*P*	*R*	*F*	*P*	*R*	*F*
Anaphylactic shock	29	22	7	32	24	28	55	35	43
Angioedema	52	32	13	40	25	30	43	26	33
Embolic and thrombotic events...	132	159	48	30	36	32	32	39	35
Haemodynamic oedema, effusions...	36	22	7	32	20	24	54	33	41
Peripheral neuropathy	31	24	13	54	42	47	96	56	71
Periorbital and eyelid disorders	39	22	16	73	41	52	77	42	54

Table 7: Qualitative evaluation of clusters corresponding to six *SMQs*: *Anaphylactic shock*; *Angioedema*; *Embolic and thrombotic events, arterial*; *Haemodynamic oedema, effusions and fluid overload*; *Peripheral neuropathy*; *Periorbital and eyelid disorders*.

Comparison of the generated clusters with the baseline (46 SMQs)

The comparison of the results generated by our method and the baseline are presented in Figure 5. The average performance of the proposed method is P=59, R=26, F=33, while the baseline average performance is P=60, R=9, F=15. With the proposed method, F-measure and recall are better than those obtained with the baseline: we gain respectively 18 and 17 points. Only precision loses one point (60 with the baseline, 59 with our method). Here again, we can observe that the performance variability across the *SMQs* is high. The general observation here is that, from the point of view of reproducing the *SMQs*, the proposed method reaches this objective better than the baseline subsumption approach usually used in the field.

Analysis of the generated clusters with an expert

Several clusters have been analyzed with an expert. We here summarize the analysis for six randomly selected *SMQs* (*Angioedema*, *Anaphylactic shock*, *Embolic and thrombotic events arterial*, *Peripheral neuropathy*, *Haemodynamic oedema, effusions and fluid overload*, and *Periorbital and eyelid disorders*). Table 7 contains information on the number of terms in these *SMQs* and in the corresponding clusters *clu*, as well as the number of common terms *com* between them. It then indicates the performance (precision P, recall R and F-measure F) when computed against the reference *SMQs Reference* and also after the analysis performed by the expert *Expert*.

We can observe similar situations across the generated clusters:

- they contain terms included in the *SMQs*, such as *Anaphylactic transfusion reaction*, *Acute respiratory failure*, *Neonatal respiratory failure* or *Anuria* for the *SMQ Anaphylactic shock*. These terms define the precision which is shown in Table 7, columns *Reference*;

- they may contain terms which are not included in the *SMQs*. These may be true false positives (such as *Solar urticaria, Urticaria thermal, Urticaria contact* for the *SMQ Angioedema*, or *Acute pulmonary oedema, Polyuria, Pulmonary renal syndrome* for the *SMQ Anaphylactic shock*), but some of these terms could also be considered for inclusion in the *SMQs*: for instance, *Injection site urticaria* and *Injection site swelling* could be included in the *SMQ Angioedema*, while *Allergic transfusion reaction, Cardiac respiratory arrest* and *Neonatal respiratory arrest* in the *SMQ Anaphylactic shock*;

Drug taken/EBGM test	HLT	SMQ	Dist
1 ACETYLSALICYLIC ACID SRT	1,66	0,474	0,326
2 ELOXATIN	0,014	0,992	1,092
3 FENTANYL CITRATE	0,189	1,143	1,88
4 FLOVENT	0,026	0,776	1,192
5 HUMULIN N	0	0,011	0,005
6 ISOSORBIDE MONONITRATE	0,065	0,806	0,941
7 JANUVIA	2,586	1,048	0,861
8 LEFLUNOMIDE	0,059	0,378	0,979
9 LISINOPRIL	0,421	0,622	0,697
10 LORAZEPAM	0,943	1,136	1,02
11 METHADONE HCL	0,816	0,895	0,894
12 METHOTREXATE SODIUM	0,529	1,915	2,138
13 NEVIRAPINE	0,345	0,94	1,02
14 PRAVACHOL	0,01	0,228	0,293
15 SOLIRIS	1,318	0,808	0,701
16 SUTENT	0,337	0,878	1,183
17 TORSEMIDE	0,56	0,915	1,135
18 VIOXX	0,103	0,174	0,171
19 ZOLPIDEM	0,146	2,005	2,345

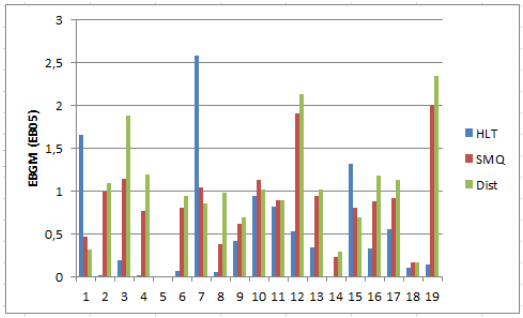


Figure 6: Signal detection for *Anaphylactic shock* safety topic. The axis x on the right graph corresponds to the 19 drugs from the table on the left.

- they may also miss relevant terms which either are not part of the ontoEIM resource, like *Anaphylactic shock* or *Circulatory collapse* (because these MedDRA terms are not aligned with SNOMED CT), or are too distant with other relevant terms within ontoEIM, such as *Anaphylactoid reaction* or *Hepatic congestion* for the *SMQ Anaphylactic shock*.

In other words, manual analysis of the clusters detected some terms which could be considered for inclusion in the *SMQs*. If we take them into account, the corrected performance of our method, indicated in the *Expert* columns in Table 7, is usually improved.

Evaluation of the generated clusters through signal detection

Thanks to this last evaluation, we analyze the impact of the generated clusters on signal detection. In Figure 6, we present the results obtained

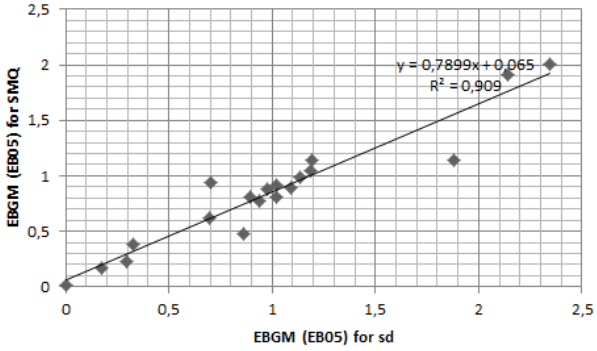


Figure 7: Correlation between the reference *SMQs* and the cluster-generated signals for the safety topic *Anaphylactic shock*.

with the EBGM method for the safety topics *Anaphylactic shock* with the corresponding *HLT* term *Anaphylactic responses*, the *SMQ* and the generated cluster. On the left handside, we indicate the 19 drugs tested and the signal detection results. The main content of this figure is represented with red, yellow and green arrows: red arrows indicate that the signal is not detected, yellow arrows indicate when the signal is correctly but feebly detected, while green arrows indicate the detection of a strong signal. For each arrow the numerical value of the signal strength is also provided. On the right handside, we can see the graphical representation of the same information. Globally, four signals (both feeble and strong) are detected with the *HLT*, five with the reference *SMQ* and nine with the generated cluster: the cluster appears to be more sensitive and efficient in this context. Figure 7 shows the correlation between the reference *SMQ* and the cluster-generated signals. The results are quite similar: the correlation is very high and close to 0.9. The small difference between them leads to a better signal detection with the generated clusters. If we look for detailed information with the cluster, we obtain two

strong signals (*Methotrexate sodium* and *Zolpidem*) and seven feeble signals (*e.g.*, *Eloxatin, Flovent, Neviparine, Sutent, Torsemide*). Several of these signals are not detected with the *SMQ* or the *HLT* (*Eloxatin, Flovent, Neviparine, Sutent, Torsemide*). Our cluster fails to return one strong (*Januvia*) and two feeble (*acetylsalicylic acid SRT* and *Torsemide*) signals. Although the evaluation of clusters against the *SMQs* shows that our methods generate smaller sets of *ADR* terms (the generated clusters display an important precision but a rather feeble recall), we can now observe that the signal detection may be well managed with these clusters. A more complete evaluation with other safety topics is ongoing.

Main factors which influence results

Several tests introduced in the methods make it possible to detect factors which have an impact on the results:

- *Semantic distances.* Among the three semantic distances applied in this study (*Rada*, *LCH* and *Zhong*), it is surprising to observe that the simplest measure *Rada*, which only relies on the counting of edges, appears to be the most efficient. It is also interesting to note that the *LCH* algorithm provides results that are very close to those of the *Rada* algorithm. As for the *Zhong* approach, its specificity (absolute depth of terms) does not seem to be relevant for the clustering of pharmacovigilance terms. Indeed, in our task, it may be important to cluster terms from low and high hierarchical levels, while the *Zhong* algorithm favors hierarchically low terms.

- *Semantic decomposition.* An important difference is observed in relation to the semantic decomposition: performance is better when we only use the *ADR* terms, without their semantic decomposition. We assume this situation is due to the incompleteness of the currently available semantic primitives: when the semantic decomposition is not complete (only one primitive is available), the semantic distance measure is biased and leads to a distorted semantic representation and to a wrong clustering.

- *Axes coefficients.* With semantic decomposition, we tested several coefficients of axes and observed that the *Disorders* axis D is to be favored because it provides important indicators for several safety topics. In other words, the following coefficients suit our purpose best: $W_{ADR} = 1$, $W_D = 2$, $W_B = 1$.

- *Clustering methods.* Among the two clustering methods tested (*Radius* and *HAC*), the *Radius* approach generally appears to provide better results. We assume this is due to the fact that the *Radius* approach generates non disjoint clusters.

- The use of *narrow* or *broad* versions of the *SMQs* has shown but a very small difference. This is a surprising result because we expected that the *narrow* version of the *SMQs* would be easier to generate.

7. Discussion

Some of the results have been discussed in the previous section. The present discussion is dedicated to the strength and limitations of the proposed

method and the obtained results, as well as their relation to existing work.

First, it should be noted that the proposed research is done within a field which is currently underexplored from the point of view of computational methods: manual and expert approaches are traditionally used there. The main input of computational methods comes from statistical tools (*i.e.*, data mining disproportionality methods), but little or no semantic methods and resources are proposed and exploited. Such deficiency makes it difficult for us to fill in the gap completely, but we hope that our study provides a good contribution to this topic. Also noteworthy is that the pharmacovigilance area is very demanding because it is closely related to the legal and industrial frameworks. Automatic methods and tools must prove to be highly reliable and efficient before they are used by the pharmacovigilance area experts.

Positive aspects of the study

General positive aspects of our study include the proposal and adaptation of a new semantic method to group the *ADR* terms which relies on an original exploitation of the semantic distance algorithms, as well as the important effort made for the evaluation of the generated results. Table 8 summarizes other positive aspects of the proposed approach. These points have different integration levels within the pharmacovigilance area: from the creation and rating of simple *ADR* term pairs, through creation of clusters with these terms and up to the impact of these clusters on the signal detection.

Correlation between the rating of the generated term pairs and human judgment. The semantic similarity between terms computed with the proposed semantic similarity approaches (*Rada* [18], *LCH* [24], and *Zhong* [20]) and the

Evaluation type	*Proposed method*	*Existing work/ Baseline*	
Correlation of term pairs rating with human judgment:			
medical doctors	0.72	0.35	
coders	0.78	0.50	
Comparison between clusters and baseline with the 46 reference SMQs:			
precision (average)	59	60	
recall (average)	26	9	
F-measure (average)	33	15	
Evaluation through signal detection process:		HLTs	SMQs
number of strong signals	2	1	1
number of weak signals	7	3	4

Table 8: Main positive aspects of the generated results: the proposed method usually outperforms existing work and the baseline (the higher the numbers the best the performance).

semantic resource ontoEIM [12] appears to be very close to human judgment: its correlation is 0.72 with medical doctors and 0.78 with medical coders. We assume that the medical coders, who are more used to manipulating and working with medical terms, and obtain a better correlation among them (0.78), also propose more correct reference annotations. The fact that we obtain up to 0.78 correlation with this evaluation set is a very good result achieved by our method. Moreover, the values of semantic distance automatically computed with our method and resource also outperform the results reported on in previous study [51]: where previous studies show 0.35 and 0.50 correlations, our approach is better by 0.37 and 0.28 against the reference annotations provided by medical doctors and coders respectively. This is also a very positive result, which may indicate that the semantic distance computed for term pairs from our term set may provide a good and precise basis for the creation of clusters of the *ADR* terms.

Comparison of the generated clusters with the baseline and the reference SMQs. Other comparative elements come from the difference we can observe on the same test set between the results provided by our method and by the baseline: our method outperforms the baseline F-measure by 18 points (we gain 17% on the baseline recall but lose 1% on precision). This indicates that our method is efficient in the pharmacovigilance context related to the creation and reproduction of *SMQs*. As explained above, we have chosen the baseline which is the most frequently used approach for the grouping of MedDRA terms: hierarchical subsumption of *PTs* through the *HLT* MedDRA level [9, 10, 52]. Hence, the baseline is an authoritative approach used for this task before the *SMQs* have been created and used. Up to now, some

experiments address the assessment and comparison between *SMQs*, *HLTs* and simple *PTs* for the detection of pharmacovigilance signals [9, 10, 52], in which the *SMQs* are not always the most efficient tool (they are often judged as too permissive). More generally, the average evaluation values of the generated clusters against the whole set of reference *SMQs* (84) are close to those obtained with the reduced set of *SMQs* (46): P=52, R=25, F=31. In both cases, the generated clusters often present a good precision while being limited by their recall. As observed above, the clusters usually correspond to different facets of the existing *SMQs*: we assume that several clusters should be considered together to guarantee a better recall for a given *SMQ*. This point has been partially studied up to now. Beyond the detection of semantic relations between terms (*i.e.*, with semantic distance algorithms), the proposed method also attempts to detect medical relations between these terms and to select those which are closely related to a given safety topic. We can observe that in this context our method is quite efficient, although the research problem addressed remains difficult.

Evaluation through signal detection. Evaluation of the generated clusters through detection of pharmacovigilance signals gives yet another indication on the suitability of the proposed method within this applied context. This evaluation is done with the FDA AERS pharmacovigilance database. Here, the assessment of the number of detected signals clearly indicates that the generated clusters allow to detect the highest number of signals (both strong and weak) by comparison with the signals detected with more traditional approaches: *HTLs* (our baseline) and *SMQs*. This is also a very positive and encouraging result.

Limitations of the proposed method and of the obtained results

Our main limitation is related to the ontoEIM resource, which suffers from the partial alignment rate with SNOMED CT: only 46% of the MedDRA PT terms are aligned through the UMLS. This means that terms which are not aligned cannot be used for the creation of clusters nor for the signal detection. To remedy this situation, we are working on the MedDRA terminology alignment: we process terms with lexical mapping and semantic categorization methods [53] and we also rely on existing work on the MedDRA terms alignment [54, 55]. With such improvements, we also hope to significantly improve the clustering results in the future.

Another limitation is due to the fact that only one expert was involved in manual analysis of the generated clusters, while ideally two or more experts should be involved. Although we believe that several kinds of the evaluation performed here allow to moderate this limitation, we plan to recruit more experts for this manual evaluation phase. In the same way, evaluation through signal detection has been performed up to now with only one safety topic *Anaphylactic shock*. A more systematic evaluation is also ongoing.

Finally, the applied clustering algorithms should also evolve. The *Radius* algorithm is a rather simple approach. We plan to test other algorithms which generate non disjoint clusters, such as those mentioned in the state-of-the-art review [40, 42, 43]. Additionally, hierarchical clustering, which appears to be quite competitive, can be used in a better way. For instance, we started experiments on the generation of hierarchically structured *SMQs* and the results proved to be encouraging [44]. We therefore plan to test whether we can generate hierarchies of clusters, for which the hierarchical

clustering should be particularly suitable.

Relation with existing work on the grouping of ADR terms

There is little existing work on the automatic creation of groupings of *ADR* terms. In addition to the MedDRA hierarchical subsumption, which we take as baseline and discussed earlier in this section, we will now try to compare our study with other usages of the ontoEIM resource. In fact, ontoEIM rather corresponds to a method than to a given resource: within the pharmacovigilance area, this method can be applied to different *ADR* terminologies, MedDRA and WHO-ART, and also to different versions of these terminologies and of SNOMED CT, according to the UMLS used version. The WHO-ART-derived resources can show up to 85.9% alignment rate with the SNOMED CT terms. Previous exploitations of this family of resources are the following:

- semantic similarity within a MedDRA-derived resource subset [16]: no comparison is done with the *SMQs*;
- semantic similarity within a WHO-ART-derived resource subset [17]: no comparison is done with the *SMQs*, but the authors provide a manual analysis and description of semantic relations within generated clusters (such as synonyms, antonyms, associated symptoms, etc. mentioned in the *Related work* section);
- hierarchical subsumption within a WHO-ART-derived resource [12]: a general method for the automatic creation of this kind of resources is presented and no comparison is done with the *SMQs*;

- hierarchical subsumption and terminological reasoning within a WHO-ART-derived resource [15]: a comparison with 24 *SMQs* is done, which shows an average sensitivity of 0.82, within the interval [0.45; 1], while the specificity is not evaluated;

- finally, a more recent study proposed to enrich a MedDRA terminology subset using a MedDRA-derived ontoEIM resource and to apply an important manual expertise [56] to reach 34,994 concepts and 157,572 definitional axioms. The resulting resource is adapted to the creation of 13 safety topics, such as *Acute renal failure*, *Agranulocytosis*, *Gastrointestinal haemorrhages*, *Peripheral neuropathy* or *Rhabdomyolysis*. Specific OWL queries have been applied and show the following average performance: P=51.1, R=63.4, F=54.

Only the last study cited above provides complete evaluation results, although these are provided for 13 safety topics to which this resource has been specifically adapted. With such extensive manual work, precision values remain comparable (51.1 against 52 obtained with our method), while recall is significantly improved thanks to the manual enrichment of the resource (63.4 against 25 obtained with our method). It is important to note that our method cannot be applied to this resource because its hierarchical structure is very poor: four MedDRA hierarchical levels (to reach the PT terms) against 14 hierarchical levels within the used version of ontoEIM. Other cited experiments cannot be compared with our study because of the different resources used or because of the lack of any evaluation. We beleive that the advantages of our method are the following: (1) it does not require a dedicated semantic resource as the publicly available alignments within

the UMLS can be used; (2) its building is easy and rapid; (3) the proposed approach is not specific to the pharmacovigilance context and can be used in other medical contexts and applications whenever the semantic relations between terms are required (information retrieval and extraction, terminology structuring, facet terminologies creation, etc.).

8. Conclusions and Perspectives

The proposed method applies the semantic distance and clustering algorithms for generating clusters of the *ADR* MedDRA terms. Several experiments have been performed for testing different factors which may influence the method performance. Among the tested semantic distance algorithms, the *Rada* approach is the most efficient. Semantic decomposition has a negative impact. The *Radius* clustering approach, which generates non disjoint clusters, is more suitable for the aimed task because terms may belong to several clusters. Evaluation against the term pairs manually rated by several experts indicates that our method provides results very close to human judgment. Evaluation related to the pharmacovigilance area shows that, although we do not generate exact content of the *SMQs*, the clusters seem to be suitable for the signal detection task. Additionally, manual evaluation by an expert indicates that the generated clusters contain relevant terms which may be missing in the *SMQs*.

The research topic addressed in our study is underexplored, which leaves room for several perspectives. For instance, current performance varies according to *SMQs* and it appears that different strategies should be used for different safety topics, while currently the same parameters of the method is

applied to all safety topics. We also plan to prepare a set of filters for performing the post-processing of the generated clusters. The proposed methods can be applied to other terminologies: we have indeed started to test their portability [57]. Testing of other clustering algorithms is also ongoing, as well as better exploitation of the hierarchical classification. Moreover, the spectral clustering [58] may be used for optimization of matrices and creation of non-disjoint clusters. We also started to apply and combine these results with those provided by Natural Language Processing: the first experiments are encouraging and we plan to strengthen this perspective [59]. Finally, we hope to have an opportunity to test the proposed clusters with the experts involved in the creation of *SMQs*.

Acknowledgements

The authors acknowledge the support from the French Agence Nationale de la Recherche (ANR) and the DGA under the Tecsan grant number ANR-11-TECS-012, and from the FP7/2007-2013 for the Innovative Medicine Initiative (IMI) under Grant Agreement number [1150004]. The authors thank Laëtitia Dupuch, Thierry Hamon, Ola Caster, Cédric Bousquet, Eric Sadou and Julien Souvignet, but the views expressed here are those of the authors only. The authors are very grateful to the anonymous reviewers for their patience, comments and help while improving the quality of the manuscript. Finally, the authors thank Marie-Aude Lefer, Louise Deléger, Gayo Diallo and Jean-Christophe Behar for the editorial assistance.

References

[1] Aagaard L, Strandell J, Melskens L, Petersen P, Holme Hansen E. Global patterns of adverse drug reactions over a decade: analyses of spontaneous reports to VigiBase. Drug Saf 2012;35(12):1171–82.

[2] Biagi C, Montanaro N, Buccellato E, Roberto G, Vaccheri A, Motola D. Underreporting in pharmacovigilance: an intervention for italian GPs (Emilia-Romagna region). Eur J Clin Pharmacol 2013;69(2):237–44.

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

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

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

[6] Fescharek R, Kübler J, Elsasser U, Frank M, Güthlein P. Medical dictionary for regulatory activities (MedDRA): Data retrieval and presentation. Int J Pharm Med 2004;18(5):259–69.

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

[8] Merrill G. The MedDRA paradox. In: AMIA Annu Symp Proc. 2008, p. 470–4.

[9] Mozzicato P. Standardised MedDRA queries: their role in signal detection. Drug Saf 2007;30(7):617–9.

[10] Pearson R, Hauben M, Goldsmith D, Gould A, Madigan D, O'Hara D, et al. Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform 2009;78(12):97–103.

[11] CIOMS . Development and rational use of standardised MedDRA queries (SMQs): Retrieving adverse drug reactions with MedDRA. Tech. Rep.; CIOMS; 2004.

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

[13] Stearns M, Price C, Spackman K, Wang A. SNOMED clinical terms: overview of the development process and project status. In: AMIA. 2001, p. 662–6.

[14] NLM . UMLS Knowledge Sources Manual. National Library of Medicine; Bethesda, Maryland; 2008. Www.nlm.nih.gov/research/umls/.

[15] Jaulent M, Alecu I. Evaluation of an ontological resource for pharmacovigilance. In: MIE. 2009, p. 522–6.

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

[17] Iavindrasana J, Bousquet C, Degoulet P, Jaulent M. Clustering WHO-ART terms using semantic distance and machine algorithms. In: AMIA. 2006, p. 369–73.

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

[19] Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network. In: CIKM. 1993, p. 67–74.

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

[21] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proceedings of Associations for Computational Linguistics. 1994, p. 133–8.

[22] Jarmasz M, Szpakowicz S. Roget's thesaurus and semantic similarity. In: Recent Advances in Natural Language Processing. 2003, p. 212–9.

[23] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research (JAIR) 1999;11:95–130.

[24] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification; chap. 4. 1998, p. 305–32.

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

[26] Lin D. An information-theoretic definition of similarity. In: International Conference on Machine Learning. Morgan Kaufmann; 1998, p. 296–304.

[27] Hirst G, St Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. 1998, p. 305–32.

[28] Steichen O, Daniel-Le Bozec C, Thieu M, Zapletal E, Jaulent M. Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus. Comput Biol Med 2006;36(7-8):768–88.

[29] Cho M, Choi J, Kim P. An efficient computational method for measuring similarity between two conceptual entities. In: WAIM 2003. 2003, p. 381–8.

[30] Yang D, Powers DMW. Measuring semantic similarity in the taxonomy of WordNet. In: Proceedings of the 28th Australasian Computer Science Conference. 2005, p. 315–22.

[31] Rosse C, Mejino J. A reference ontology for biomedical informatics: the foundational model of anatomy. J Biomed Inform 2003;36(6):478–500.

[32] Fellbaum C. A semantic network of English: the mother of all WordNets. Computers and Humanities EuroWordNet: a multilingual database with lexical semantic network 1998;32(2-3):209–20.

[33] Lord P, Stevens R, Brass A, Goble C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19(10):1275–83.

[34] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Eighteenth International Joint Conference on Artificial Intelligence. 2003, p. 805–10.

[35] McInnes B, Pedersen T, Pakhomov S. UMLS-interface and UMLS-similarity : Open source software for measuring paths and semantic similarity. In: AMIA. 2009, p. 431–5.

[36] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the umls. Journal of Biomedical Informatics 2004;37:77–85.

[37] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967, p. 281–97.

[38] Kaufman L, Rousseeuw P. Clustering by means of medoids. 1987, p. 405–16.

[39] Bezdek J. Pattern Recognition with Fuzzy Objective Function Algoritms. New York: Plenum Press; 1981.

[40] Krishnapuram R, Joshi A, Nasraoui O, Yi L. Low complexity fuzzy relational clustering algorithms for web mining. In: IEEE Trans. Fuzzy System. 2001, p. 595–607.

[41] Lelu A. Modles neuronaux pour lanalyse de donnes documentaires et textuelles. Phd thesis; Universite de Paris VI; Paris, France; 1993.

[42] Cleuziou G, Martin L, Vrain C. PoBOC: an overlapping clustering algorithm. application to rule-based classification and textual data. In: ECAI. 2004, p. 440–4.

[43] Cleuziou G. OKM: une extension des k-moyennes pour la recherche de classes recouvrantes. In: EGC. 2007, p. 691–702.

[44] Dupuch M, Bousquet C, Grabar N. Automatic creation and refinement of the clusters of pharmacovigilance terms. In: ACM IHI. 2012, p. 181–90.

[45] Johnson S. Hierarchical clustering schemes. Psychometrika 1967;32:241–54.

[46] Kaufman L, Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley; 1990.

[47] Zhang T, Ramakrishnan R, Livny M. Birch: an efficient data clustering method for very large databases. In: ACM SIGMOD. 1996, p. 103–14.

[48] Guha S, Rastogi R, Shim K. Cure: an efficient clustering algorithm for large databases. In: ACM SIGMOD. 1998, p. 73–84.

[49] Dupuch M, Lerch M, Jamet A, Jaulent M, Fescharek R, Grabar N. Grouping pharmacovigilance terms with semantic distance. In: Proc of MIE. 2011, p. 794–8.

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

[51] Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(3):288–99.

[52] Yuen N, Fram D, Vanderwall D, Almenoff J. Do standardized MedDRA queries add value to safety data mining? In: ICPE 2008. 2008, p. 1–2.

[53] Mougin F, Dupuch M, Grabar N. Improving the mapping between MedDRA and SNOMED CT. In: AIME. 2011, p. 220–4. Springer LNAI 6747.

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

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

[56] Declerck G, Bousquet C, Jaulent M. Automatic generation of MedDRA terms groupings using an ontology. In: MIE. 2012, p. 73–7.

[57] Homo J, Dupuch L, Benbrahim A, Grabar N, Dupuch M. Customization of biomedical terminologies. In: Stud Health Technol Inform. 2012, p. 153–8.

[58] Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing 2007;17(4):395–416.

[59] Dupuch M, Dupuch L, Hamon T, Grabar N. Semantic distance and terminology structuring methods for the detection of semantically close terms. In: NAACL BIONLP. 2012, p. 109–17.