# Grouping the pharmacovigilance terms with a hybrid approach

Marie DUPUCH[a], Laëtitia DUPUCH[b], Amandine PERINET[c],
Thierry HAMON[c] and Natalia GRABAR[a]
[a]*CNRS UMR8163, Université Lille 1&3, France;*
[b]*Université Toulouse III Paul Sabatier, France;*
[c]*LIM&BIO (EA3969)Université Paris 13, Bobigny, France*

**Abstract.** Pharmacovigilance is the activity related to the collection, analysis and prevention of adverse drug reactions (ADRs) induced by drugs. It leads to the safety survey of pharmaceutical products. The pharmacovigilance process benefits from the traditional statistical approaches and also from the qualitative information on semantic relations between close ADR terms, such as SMQs or hierarchical levels of MedDRA. In this work, our objective is to detect the semantic relatedness between the ADR MedDRA terms. To achieve this, we combine two approaches: semantic similarity algorithms computed within structured resources and terminology structuring methods applied to a raw list of the MedDRA terms. We compare these methods between them and study their differences and complementarity. The results are evaluated against the gold standard manually compiled within the pharmacovigilance area and also with an expert. The combination of the methods leads to an improved recall.

**Keywords.** Adverse Drug Reaction Reporting Systems; Algorithms; Artificial Intelligence; Automatic Data Processing; Vocabulary, Controlled

## 1. Introduction

Pharmacovigilance is the activity related to the collection, analysis and prevention of adverse drug reactions (ADRs) induced by drugs. Detection of a new serious ADR may modify the conditions of the use of the product, to reduce its use or even to withdraw the product from the market. Safety signal detection – the detection of previously unexpected potentially causal associations between drugs and ADRs – depends on the quality and specific features of the ADR coding. Currently, the ADRs are coded with the MedDRA terminology [1] (Medical Dictionary for Drug Regulatory Activities). For the analysis of these databases and the signal detection, traditional pharmacovigilance methods [2-3] are exploited. They are currently supplemented by statistical algorithms [4-5]. To improve the signal detection, these methods benefit from groupings of related ADR terms [6], which are especially relevant because the structure of MedDRA is very fine-grained and closely related terms can be spread in this terminology: the use of very specific terms for coding ADRs may cause a dilution of signals [7]. In that purpose, SMQs (Standardized MedDRA Queries) have been created. They gather MedDRA terms specific to a given medical condition. The SMQs are defined by groups of experts through a manual study of the MedDRA's structure and the scientific literature [8]. It is a long and meticulous task. Now there are 84 SMQs that cover several important medical conditions, as for instance *Acute renal failure*, *Agranulocytosis*, *Angioedema*, etc. But several other SMQs are still to be defined. The terms present in a SMQ belong to different SOCs (System Class Organ) of MedDRA. Within the 84 existing SMQs, the variety of SOCs varies between 4 and 25 (the full number of SOCs being 32), while the average is of 8,26 SOCs by SMQ. In addition, a same term can belong to more than one SMQ. Indeed, the ADRs can appear in relation to different medical conditions. These observations show that the recruitment of the terms for the SMQs follows a very precise medical logic and does not especially respect the MedDRA structure into SOCs. Because the creation of the SMQs is a long and meticulous process, we propose an automatic method to assist this process. There are very few existing works: grouping of the ADR terms through hierarchical subsumption [9-10] or semantic distance [11-12], or extension of Pubmed queries for the pharmacovigilance [13]. Only one previous work [10] has been partially evaluated against the SMQs. In our previous work [14], we also exploited the semantic similarity measures and evaluated the obtained groupings against the SMQs. The precision is usually high, although the recall remains low. In this work, we propose to improve these results and to combine semantic similarity algorithms with a Natural Language Processing (NLP) method dedicated to the terminology structuring. We aim at the detection of synonym and hierarchical relations. We analyze the results provided by these two approaches and evaluate them against the SMQs and with an expert.

## 2. Material and Methods

**Material.** We rely on material issued from MedDRA (ontoEIM resource and SMQs) and on linguistic resources. ontoEIM [9] gathers and structures the ADR terms. It has been created thanks to the projection of MedDRA on SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [15] through the UMLS (Unified Medical Language System) [16]. Although only 46% of MedDRA terms are currently available within ontoEIM, their structure is improved (up to 14 hierarchical levels against only five levels in MedDRA) and they receive formal definitions (the meaning of ADRs is decomposed on up to four axes: morphology, topography, causality and expression). Linguistic resources provide three sets of synonyms: 1) medical synonyms extracted from the UMLS (n=228,542); 2) medical synonyms acquired from three biomedical terminologies [17] (n=28,691); 3) general language synonyms provided by WordNet [18] (n=45,782). We exploit the 84 existing SMQs as the reference data.

**Methods.** The proposed method combines two approaches and consists of four main steps (figure 1): 1) the computing of the semantic similarity between the MedDRA terms and their clustering, 2) the application of the terminology structuring approach to acquire semantic relations within a raw list of the MedDRA terms and their clustering, 3) the merging of these two sets of clusters, 4) the evaluation of these clusters.

We exploit three algorithms to compute the similarity between two terms *t1* and *t2* (two semantic distances: *Rada* [19] and *Zhong* [20] and one semantic similarity: Leacock and Chodorow *LCH* [21]) and two clustering methods (hierarchical ascendant classification performed with with R project [22] with which we tested numbers of classes included within the interval [100; 7,000] and a *R* Radius approach in which every MedDRA term is considered as a possible center of a cluster and its closest terms are clustered with it). A detailed presentation of this method can be found in [14]. It generates clusters of terms of which we exploit those obtained with the radius approach because the clusters are non disjoint and respect better the particularity of the SMQs.

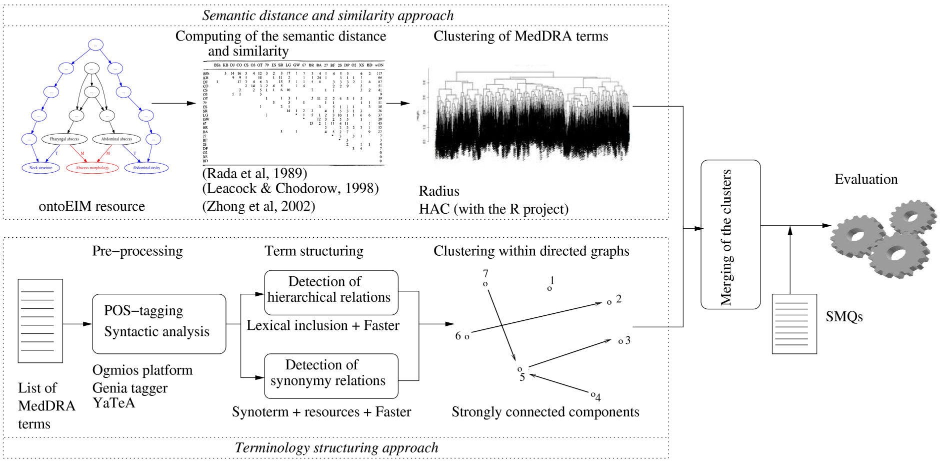
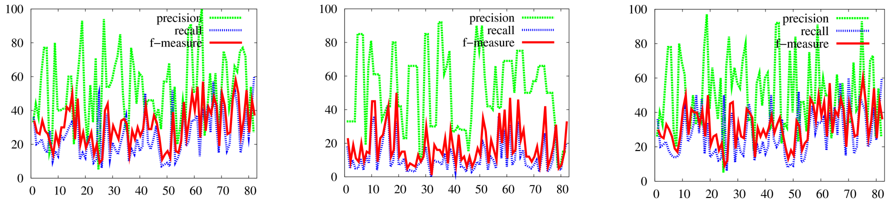

Semantic distance and similarity approach
Computing of the semantic distance and similarity
Clustering of MedDRA terms
(Rada et al, 1989)
(Leacock & Chodorow, 1998)
(Zhong et al, 2002)
ontoEIM resource
Radius
HAC (with the R project)
Merging of the clusters
Evaluation
SMQs
Pre−processing
Term structuring
Clustering within directed graphs
List of MedDRA terms
POS−tagging
Syntactic analysis
Ogmios platform
Genia tagger
YaTeA
Detection of hierarchical relations
Lexical inclusion + Faster
Detection of synonymy relations
Synoterm + resources + Faster
7
1
2
6
3
5
4
Strongly connected components
Terminology structuring approach

***Figure 1: General schema of the hybrid approach***

The terminology structuring methods are applied to a raw list of 18,209 MedDRA ADRs. They are processed with the POS-tagger Genia [23] and the syntactic parser YATEA [24]. We detect the hierarchical relations through the lexical inclusions [25] (if one term is lexically included in another term, there is a hierarchical relation between them: the short term is the parent term and the long term is the child term). The synonymy relations are detected through their compositionality [26]. Finally, the detection of morpho-syntactic variants with Faster [27] provides hierarchical and synonymy relations according to the transformation rules. The sets of terms related through hierarchical relations are considered as directed graphs, which are partitioned into strongly connected components to obtain the clusters. To improve the coverage of the clusters, we add the synonyms: if a term has a synonymy relation with the term from a cluster then this term is also included in this cluster. For the evaluation we give a judgment about: (1) the correctness of the generated relations, (2) their relevance to the creation of the SMQs evaluated against the reference data, (3) their relevance to the creation of the SMQs through a manual evaluation by an expert. The evaluation of the clusters is performed with the three measures: precision P, recall R and f-measure F.

## 3. Results and Discussion

The MedDRA terms from the ontoEIM have been processed with semantic measures and clustered. The best thresholds with the *Radius* clustering, empirically defined, are 2 for *Rada*, 4.10 for *LCH* and 0.02 for *Zhong*. The best parameters are the *Rada* distance, no formal definitions and Radius clustering approach. The raw list of the MedDRA terms has also been processed through the NLP and terminology structuring methods. The best experience is when the lexical inclusions are augmented by Faster and by compositionally computed synonyms. A manual analysis of the hierarchical relations indicates that these relations are always correct: the constraint involved through the syntactic analysis guarantees a correct generation. The clusters provided by these methods have been merged into 2,906 clusters with 1 to 563 terms per cluster (the mean is 25 terms/cluster).



(a) Semantic distance (b) Terminology structuring (c) Merging of the two approaches

***Figure 2: Results (precision, recall and f-measure) for the semantic distance and terminology structuring approaches and for the merging of the two approaches***

Figure 2 provides a quantitative evaluation of the clusters: semantic distance (figure 2(a)), terminology structuring (figure 2(b)), and merging of these two sets (figure 2(c)). We can observe that there is a great variability among the SMQs and the two approaches but that the precision (green lines) is systematically high while the recall (blue lines) is rather low. The positive result is that these approaches are indeed complementary: their merging slightly increases the performance and especially the recall. An analysis of the clusters generated with terminology structuring shows that: (1) hierarchical relations correspond to the core of the clusters (up to 96% of the involved terms) and show 69% precision. Only three clusters do not contain hierarchical relations; (2) Faster relations are involved in 50% of clusters and show precision between 75 and 85%; (3) one third of the clusters contains synonymy relations, their precision varies between 55 and 69%.

We also performed a detailed analysis of the noise in several clusters with an expert. This analysis indicates that across the clusters we have similar situations because they may contain false positives which are non relevant to a given medical condition. Further to this analysis, we also observed that the SMQs may contain very general and non relevant terms or, on contrary, they may miss relevant terms as it was indicated in a previous work [28]. Notice that with our approach we have found some of these missing terms. The corrected performances of the generated clusters are improved by several points. Our experiences indicate that the proposed automatic approaches may provide a useful basis for the creation of SMQs, especially because they systematically collect all the relevant terms which satisfy the given algorithmic conditions.

## 4. Conclusion and Perspectives

We have applied two different approaches for the clustering of pharmacovigilance terms. We performed a comparison of the results obtained with these two approaches and analyzed their complementarity. Although the automatic creation of the SMQs is a difficult task, our results seem to indicate that the automatic methods may be used as a basis for the creation of new SMQs. The precision of the clusters is often satisfactory, while their merging leads to the improvement of their completeness. Future studies will lead to the identification of other parameters 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 methods to all the SMQs. Different filters (i.e., lexical and hierarchical) will be tested to clean up the results and to remove the true false positive terms. Besides, the obtained clusters will also be evaluated through their impact on the exploring of the pharmacovigilance databases. The improvement of the drug safety survey, as the first results non presented in this paper suggest, is the main practical impact of our work.

## References

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

[2] Almenoff JS, Tonning JM, Gould AL & al. (2005). Perspectives on the use of data mining in pharmacovigilance. Pharmacoepidemiol Drug Saf., **28**, 981-1007.

[3] Bailey S, Singh A, Azadian R & al. (2010). Prospective data mining of six products in the US FDA Adverse Event Reporting System: disposition of events identified and impact on product safety profiles. Pharmacoepidemiol Drug Saf., **33**(2), 139–46.

[4] Bate A, Lindquist M, Edwards I & al. (1998). A bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol, **54**(4), 315–21.

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

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

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

[8] CIOMS (August 2004). Development and Rational Use of Standardised MedDRA Queries (SMQs): Retrieving Adverse Drug Reactions with MedDRA. Report of the CIOMS Working Group, CIOMS.

[9] Alecu I, Bousquet C & Jaulent MC. (2008). A case report: using SNOMED CT for grouping adverse drug reactions terms. BMC Med Inform Decis Mak, **8**(S1), 4.

[10] Jaulent MC & Alecu I. Evaluation of an ontological resource for pharmacovigilance. In Stud Health Technol Inform, pages 522--6, 2009

[11] Bousquet C, Henegar C, Lillo-Le Louët A & al.. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform*, 74(7-8):563--71, 2005

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

[13] Delamarre D, Lillo-Le Louët A, Guillot L & al. Documentation in pharmacovigilance: using an ontology to extend and normalize Pubmed queries. Stud Health Technol Inform2010: 518-22

[14] Dupuch M, Bousquet C & Grabar N. Automatic creation and refinement of the clusters of pharmacovigilance terms. In ACM IHI 2012. To appear

[15] Stearns M, Price C, Spackman K & Wang A. (2001). SNOMED clinical terms: overview of the development process and project status. In AMIA, p. 662–666.

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

[17] Grabar N & Hamon T. (2010). Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In MEDINFO, p. 1015–9.

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

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

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

[21] Leacock C. & Chodorow M. (1998). Combining local context and WordNet similarity for word sense identification. MIT Press, p. 305–32.

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

[23] Tsuruoka Y, Tateishi Y, Kim JD & al. (2005). Developing a robust part-of-speech tagger for biomedical text. In LNCS, p. 7746:382–92.

[24] Aubin S & Hamon T. (2006). Improving term extraction with terminological resources. In FinTAL, number 4139 in LNAI, p. 380–87.

[25] Kleiber G & Tamba I. (1990). L'hyperonymie revisitée: inclusion et hiérarchie. Langages, 98: 7–32.

[26] Partee BH. (1984). Compositionality. In Landman F. & Veltman F., editor, Varieties of formal semantics.

[27] Jacquemin C. (1996). A symbolic and surgical acquisition of terms through variation. In Connectionist, statistical and symbolic Approaches to Learning for Natural Language Processing, p. 425–38.

[28] Pearson R, Hauben M, Goldsmith D & al.(2009). Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform, **78**(12), 97–103.