Journal of Biomedical Informatics 54 (2015) 174–185

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

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

Marie Dupuch [a,b,c,*], Natalia Grabar [a]

[a] CNRS UMR 8163 STL; Université Lille 1&3, F-59653 Villeneuve d'Ascq, France
[b] INSERM, U872, Paris F-75006, France
[c] Viseo-Objet Direct, 4, Avenue Doyen Louis Weil, F-38000 Grenoble, France

ARTICLE INFO

Article history:
Received 5 October 2014
Accepted 13 November 2014
Available online 4 February 2015

Keywords:
Pharmacovigilance
Terminology
Clustering
Semantic distance and similarity
MedDRA
SMQs

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, 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 performance 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.

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

* Corresponding author at: CNRS UMR 8163 STL; Université Lille 1&3, F-59653 Villeneuve d'Ascq, France. Fax: +33 3 20 41 67 14.
E-mail address: dupuchm@hotmail.fr (M. Dupuch).

http://dx.doi.org/10.1016/j.jbi.2014.11.007
1532-0464/

necessary during the signal detection[1] [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 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*, etc.). 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[2] [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 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.

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

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*	1688
PT	Preferred Terms	*Bradycardia*	18,209
LLT	Lowest Level Terms	*Bradycardia*, *Bradycardiac tendency*, *Reflex bradycardia*, etc.	66,587
Total			86,842

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, etc.) and/or whether they are path-based or node-based. In Table 2, we indicate the most frequently used semantic distance measures. Measures from the first series [18–22] 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 [23–26] 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 [27–30] is that they use not only hierarchical relations, but also other types of relations (such as *treatment of*, *causes*, *finding site of*, and *associated morphology of*). 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

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

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