OntoADR a Semantic Resource Describing Adverse Drug Reactions to Support Searching, Coding, and Information Retrieval

Julien Souvignet, Gunnar Declerck, Hadyl Asfari, Marie-Christine Jaulent, Cédric Bousquet

OntoADR a Semantic Resource Describing Adverse Drug Reactions to Support Searching, Coding, and Information Retrieval

Authors

Julien Souvignet (a,b), Gunnar Declerck (c), Hadyl Asfari (a), Marie-Christine Jaulent (a), and Cédric Bousquet (a,b)

Affiliations

a: INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France

b: SSPIM, CHU University Hospital of Saint Etienne, France

c: Sorbonne Universités, Université de technologie de Compiègne, EA 2223 Costech (Connaissance, Organisation et Systèmes Techniques), Centre Pierre Guillaumat - CS 60 319 - 60 203 Compiègne cedex

Corresponding Author

Cédric Bousquet

Service de Santé Publique et de l'Information Médicale

Bâtiment CIM 42 - Hôpital Nord

Chemin de la Marandière

42270 Saint Priest-en-Jarez

+ 33 (0)4 77 12 79 74

cedric.bousquet@chu-st-etienne.fr

Abstract

Introduction: Efficient searching and coding in databases that use terminological resources requires that they support efficient data retrieval. The Medical Dictionary for Regulatory Activities (MedDRA) is a reference terminology for several countries and organizations to code adverse drug reactions (ADRs) for pharmacovigilance. Ontologies that are available in the medical domain provide several advantages such as reasoning to improve data retrieval. The field of pharmacovigilance does not yet benefit from a fully operational ontology to formally represent the MedDRA terms. Our objective was to build a semantic resource based on formal description logic to improve MedDRA term retrieval and aid the generation of on-demand custom groupings by appropriately and efficiently selecting terms: OntoADR.

Methods: The method consists of the following steps: 1) mapping between MedDRA terms and SNOMED-CT, 2) generation of semantic definitions using semi-automatic methods, 3) storage of the resource and 4) manual curation by pharmacovigilance experts.

Results: We built a semantic resource for ADRs enabling a new type of semantics-based term search. OntoADR adds new search capabilities relative to previous approaches, overcoming the usual limitations of computation using lightweight description logic, such as the intractability of unions or negation queries, bringing it closer to user needs. Our automated approach for defining MedDRA terms enabled the association of at least one defining relationship with 67% of preferred terms. The curation work performed on our sample showed an error level of 14% for this automated approach. We tested OntoADR in practice, which allowed us to build custom groupings for several medical topics of interest.

Discussion: The methods we describe in this article could be adapted and extended to other terminologies which do not benefit from a formal semantic representation, thus enabling better data retrieval performance. Our custom groupings of MedDRA terms were used while performing signal detection, which suggests that the graphical user interface we are currently implementing to process OntoADR could be usefully integrated into specialized pharmacovigilance software that rely on MedDRA.

Keywords

Knowledge Representation; Pharmacovigilance; Biological Ontologies; Terminological Reasoning; Data Retrieval

1 Introduction

Efficient searching and coding in pharmacovigilance databases using terminological resources requires that they support efficient data retrieval [01]. The increasing number of terms in terminological resources necessitates the ability to sort out terms that are relevant for a given purpose [02]. For example, retrieving data on specific safety issues among millions of case reports in pharmacovigilance databases can only succeed if appropriate medical terms are used.

MedDRA (Medical Dictionary for Regulatory Activities) is the reference terminology used by regulatory authorities and the pharmaceutical industry to code adverse drug reactions (ADR) in pharmacovigilance case reports [03]. Some authors have argued for replacing terminologies such as MedDRA with formal terminological systems to support improved processing of clinical data in modern health information systems [04]. Indeed, classical terminologies lack formal definitions, and thus need to be formalized and semantically represented. These terminologies may then benefit from the properties of formal knowledge representations to enhance searching and coding in current health information systems. The explicit representation of the meaning of terms allows computational processing and reasoning algorithms that support enhanced data retrieval [02].

We have been conducting a research program over the last several years that aims to demonstrate the benefits of adding formal definitions to MedDRA terms [05,06]. We are currently developing such an approach where concepts of MedDRA terminology are formally defined in a semantic resource of ADRs, called *OntoADR*, based on mapping with SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms).

A possible alternative to OntoADR is OAE (Ontology of adverse events). OAE takes into account 2,300 MedDRA terms, but does not provide definitions for more than 20,000 MedDRA terms that may be used for coding ADRs in pharmacovigilance databases. This is because OAE only considers 'pathological bodily process' descendants (corresponding to both Morphology and Definitional Manifestation in SNOMED-CT). MedDRA terms corresponding to investigation results (e.g. increased creatinine) or medical procedures (e.g. dialysis) are thus excluded, whereas we know that coding "Increased creatinine" or "dialysis" generally implies kidney failure. In daily routine, some users will use signs, symptoms, or investigations rather than pathological bodily processes. The objective of OAE is to describe

adverse events based on a solid ontological basis. One of our major goals when creating OntoADR was to take this a step further by identifying all MedDRA terms (investigations, procedures, etc.) that are potentially associated with a clinical condition.

In a previous study we proposed to ''ontologize'' MedDRA and described the general process for formalizing MedDRA to support semantic reasoning [06]. Our attempt consisted of the following methodological steps: (a) Automatic one to one mapping between MedDRA terms and SNOMED-CT concepts; (b) Manual validation of this mapping by knowledge engineers and pharmacovigilance experts; (c) Automatic conversion of the MedDRA hierarchy into a subsumption tree; (d) Semi-automatic completion by syntactic analysis of MedDRA labels; (e) Manual completion of a subset of MedDRA; and (f) Implementation of formal definitions in an OWL file. We encountered some limitations when we tested MedDRA search with this first version of OntoADR.

In step (a), we often observed that a single MedDRA term mapped to several SNOMED-CT concepts although they have different meanings. For example, the MedDRA term "Spondylitis" mapped in UMLS to the SNOMED-CT concepts "Undifferentiated spondylitis", "Inflammatory spondylopathy" and "Spondylitis". We had to map MedDRA terms to a single SNOMED-CT concept, that we manually selected from the ones proposed as synonyms in UMLS (in this case "Spondylitis"), to ensure that each MedDRA term had a unique definition. This strategy presented two major drawbacks. First, it required time-consuming expert validation to make this choice in step (b). Second, the SNOMED-CT definition for a concept can sometimes be incomplete or even empty. Selecting such an undefined concept (whereas another choice was defined) led to a useless mapping (in terms of value-added knowledge).

In step (c), the direct conversion of the MedDRA hierarchy into a subsumption tree resulted in semantic inconsistencies. Reasoners raised thousands of logical errors when applied to the description logic of MedDRA terms. Indeed, the MedDRA hierarchy is far from a subsumption tree. For example, groups of symptoms (HLGT or HLT) are placed under the categories of the disorders they refer to. This immediately raises inconsistency problems: a symptom of disease X is not a kind of disease X. The same applies to MedDRA terms that refer to complications. Other inadequate parent-child relationships are spread throughout MedDRA (e.g. "sudden death" child of "cardiac arrhythmias").

In Step (d) OntoADR uses EL++ [07,08] a lightweight description logic that allows polynomial time inferencing for reasoning tasks. EL++ does not allow negation which is a major drawback [09]. Rector and Brandt suggested that SNOMED should be described using OWL DL to overcome this limitation. However, reasoning on large OWL-DL ontologies proved to be "very slow and demand too much memory to operate on large KBs, if they work at all." [10]. Therefore, most authors still rely on EL++ [11,12] when performing semantic queries on large ontologies.

This article presents our new strategy for building the OntoADR resource. Our new approach simplifies MedDRA formalization with only 4 steps: (1) Automatic one-to-n mapping between MedDRA terms and SNOMED-CT concepts using multiple sources and recycling previous step (a); (2) Merging of semantic definitions from all semi-automatic methods, including some results from previous steps (b, d and e); (3) Implementation of formal definitions using a relational database representation (4) Manual curation by pharmacovigilance experts. This approach introduces significant improvements: revisions of MedDRA terms definitions,, removal of inconsistencies in reasoning, and implementation of OntoADR in a database that allows computation of negation and disjunction in limited computational time. We also aim to demonstrate that using a combination of semantic reasoning and set theory groupings can lead to advanced MedDRA information retrieval.

2 Methods

We performed alternative steps to formalize MedDRA that overcame previous observed limitations and enabled us to more precisely describe each term with semantic definitions.

2.1 Step 1: One-to-n mapping between MedDRA terms and SNOMED-CT concepts using multiple sources

We used the UMLS (Unified Medical Language System) meta-thesaurus to align MedDRA terms with SNOMED-CT concepts, where concepts are organized within a semantic network, developed by the NLM (U.S. National Library of Medicine) [13]. This network maps terms from multiple controlled vocabularies to unique UMLS concepts defined by their concept

unique identifier (CUI). As UMLS includes both SNOMED-CT and MedDRA concepts, we could extract synonymous mapped concepts (same CUI) from these terminologies.

The MedDRA hierarchy is organized in five levels. The first hierarchical level is based on the anatomical location: system organ class (SOC) e.g. "cardiac disorders" or "eye disorders". Level 2 is comprised of high level group terms (HLGT) e.g. "heart failures" or "ocular neoplasms"; followed by level 3: high level terms (HLT); level 4: preferred terms (PT); and finally level 5: low level terms (LLT). MedDRA also proposes Standardized MedDRA Queries (SMQ) that group PTs from different SOC, HLGT and/or HLT (e.g. Anaphylaxis), but these are provided in a limited number.

We converted MedDRA version 18, SNOMED-CT version March 2015, and UMLS version 2014AB files into a relational database. MedDRA term information (id, label, level, and hierarchy) was extracted using documented *.ASC files. SNOMED-CT concepts were recovered directly from the Concepts_Core_INT file (Release Format 1), which includes the id and labels, while the hierarchy and semantic properties were retrieved from the Relationships_Core_INT file using the identifier 116680003 for the 'Is a' relation. For all other required SNOMED-CT properties of OntoADR (Table 2), we identified the corresponding SNOMED-CT concepts. We then stored information about the range of each property (e.g. the Finding Site attribute is limited to the range of "Anatomical or acquired body structure", identifier 442083009).

We extracted UMLS mappings between MedDRA and SNOMED-CT in two ways. We first loaded UMLS tables using the MRCONSO.RRF file to extract the CUI associations and ids, for both MedDRA (code: "MDR") and SNOMED-CT (code: "SNOMEDCT"). We also checked the MRREL.RRF file which contains synonymy information in field 4 (code "SY"), which allowed us to extract a few additional CUI associations. We created a table dedicated to UMLS mappings between MedDRA and SNOMED-CT from the two extractions (same CUI and synonyms). For each mapping, we extracted the definition of the SNOMED-CT concept, and added it to the associated MedDRA term. When a MedDRA term was associated with several SNOMED terms (in case of 1-to-n mappings), all properties were extracted and added to the MedDRA term definition.

We also used other mapping resources such as Nadkarni & Darer's propositions of mapping between MedDRA and SNOMED-CT [14]. They defined 786 PTs with no SNOMED-

CT mapping by attempting to map them (with software assistance) via one-to-one or one-to-n mappings. They proposed a majority of compositional mappings: we had to convert them into semantic definitions. For example, if the mappings corresponded to a morphologic abnormality link; we used the "Associated Morphology" attribute. The same method applied to other properties. For example, "Abdominal Sepsis" mapped with "Abdominal structure" was transformed to "Abdominal Sepsis" Finding Site "Abdominal structure".

2.2 Step 2: Merging of semantic definitions using semi-automatic methods

We implemented a simple algorithm for automatic creation of properties from the MedDRA label. E.g., when the algorithm detects a given string Sx (hemorrhage, perforation, etc.) in a MedDRA label, it automatically adds a corresponding Px property (manifestation, morphology, etc.) to the definition.

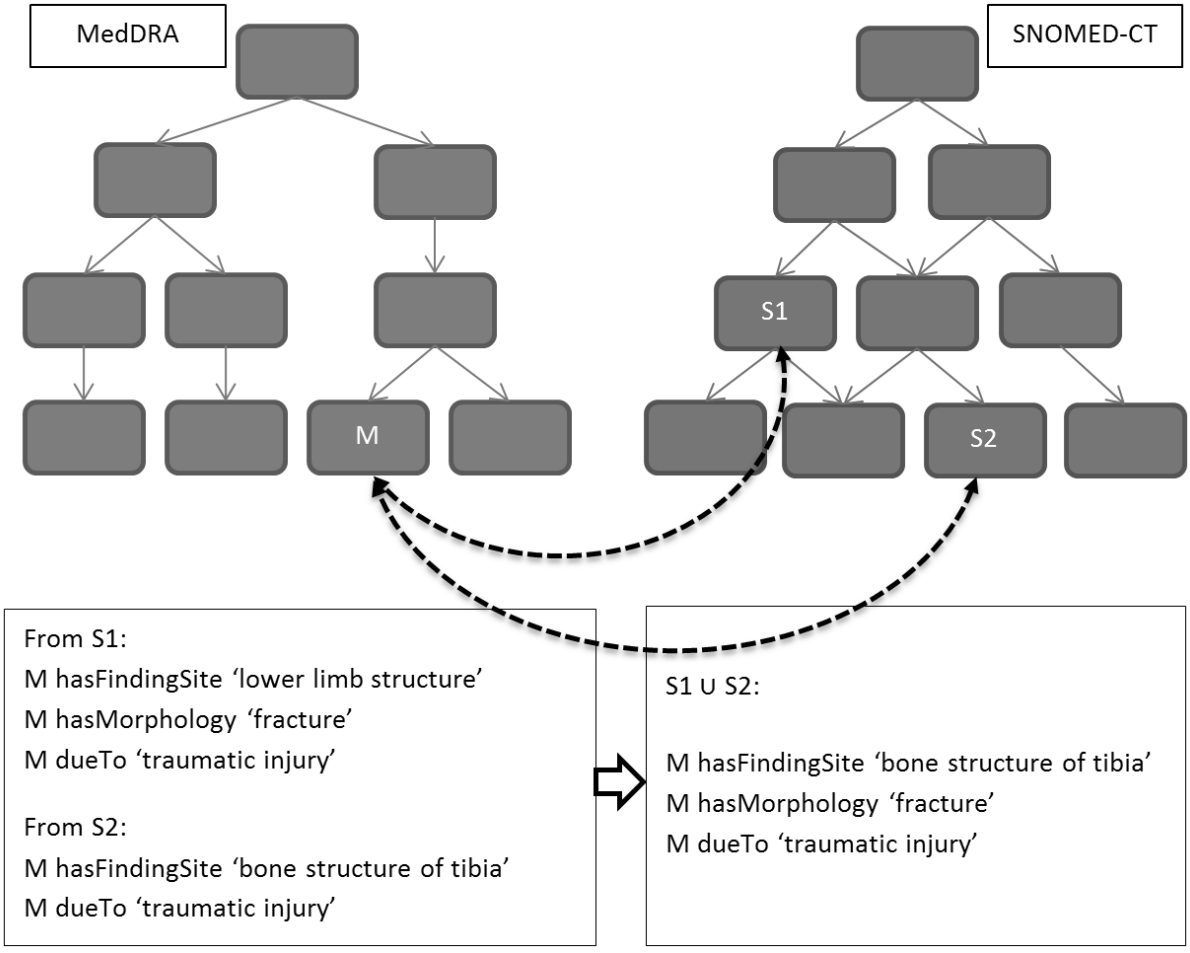
Combining MedDRA and SNOMED-CT related information introduced unnecessary relations. We developed several algorithms that performed filtering tasks to clean redundant data. 1) A first algorithm replaced or removed relations with inactive concepts from SNOMED-CT using 'ConceptStatus' attribute. Inactivated concepts are old concepts used in previous versions that have been replaced by another more accurate concept (using "same as" or "may be a" relation in SNOMED-CT). 2) A second algorithm removed duplicated relationships resulting from multiple mappings between MedDRA and SNOMED-CT, especially inferred relationships: if *a* is R-related to *b* and *a* is R-related to *c* and *c* is subsumed by *b* then knowing that *a* is R-related to *b* is implicit (it can be inferred by reasoning) and has no value for us. 3) A third algorithm removed all non-informative relationships, for example those binding the highest level element of an ontological range. Axioms such as hasSeverity some Severities (272141005), hasEpisodicity some Episodicities (288526004), or hasClinicalCourse some Courses (288524001) have no value in terms of reasoning.

Contrary to our previous approach, definitions of parents in MedDRA were no longer inherited by their children to prevent experienced inconsistencies [06].

Figure 1 presents a case were MedDRA concept M is mapped with both SNOMED-CT concepts S1 and S2. Our filters removed duplicated relationships (e.g. dueTo 'traumatic injury') and redundant properties (e.g. hasFindingSite 'lower limb structure'). The 'bone

structure of the tibia' is located in the 'lower limb structure'. If an adverse event is described with a finding site in the tibia, we already know that it is in the lower limb. There is thus no need to keep this redundant knowledge.

Figure 1: Illustration of filtering/inferring process in OntoADR



2.3 Step 3: Implementation of formal definitions using database representation

We used EL++ description logic for defining ADRs in OntoADR because we use SNOMED-CT, but were consequently limited by the impossibility of performing negation and disjunction queries. In practice, MedDRA groupings are intended to address a broad coverage and include diseases, diagnoses, signs or symptoms, indications, etc. MedDRA users have varied profiles and each will remove the terms that do not suit him. For example, users will remove all congenital diseases or terms related to pathogenic agents when searching for ADRs.

We relied on set theory and binary operations to allow users to search in OntoADR using disjunction and negation. Description logic queries yield a result in the form of an inferred hierarchy, from which entities can be extracted to create a flat list of entities (e.g. MedDRA PTs). Any set of elements can be defined in extension (by naming or designating each individual which is part of it) or in intension, by a description (specification of a number of predicates) that defines the set. Here, the intension is the query (e.g. hasFindingSite 'KidneyStructure') and the extension is the set of concepts matching the query (e.g. {renal failure, pyelonephritis, etc.}). We decided to work on the extension (set of MedDRA terms) rather than on the intension (description logic query)). Computation with the DL operators, negation and disjunction can then be replaced by set operators, difference, and union.

Negations can be calculated using the "set difference" operator in Structured Query Language ("NOT IN"), and set union can be handled with the "UNION" operator [15]. We decided to keep formal definitions in our relational database [16]. Taking into account the specific characteristics of OntoADR, we implemented optimizations such as caching [17] to manage numerous hierarchical processing that may introduce performance issues.

2.4 Step 4: Manual curation by a pharmacovigilance expert

All formal definitions were generated through automated methods and thus required validation by an expert. The curation task by a pharmacovigilance expert ("curator") consisted of manually reviewing the generated semantic definitions of MedDRA terms. Several limits in the formal definitions of ADRs were observed, especially partial definitions, fluctuating granularity, or even incorrect axioms.

Performing curation on the OWL version of OntoADR was difficult because there was no dedicated software for curation and maintenance of ontologies. We developed the Ci4SeR tool for this task [18].

We decided to limit curation work to the PT level. This is the "preferred" level for major pharmacovigilance purposes, such as signal detection. Definitions associated with parents are no longer automatically asserted to the definition of their children, as MedDRA hierarchy is ignored, but are suggested to the curator who can accept or reject them.

We performed manual curation of approximately 2000 MedDRA terms in 12 months [18]. We focused on high value-added terms for pharmacovigilance. These terms were selected on the basis of a ranked list of 23 first importance adverse drug events based on a multi-source review proposed by Trifirò et al. [19]. To identify which MedDRA terms are related to these topics, we selected their closest SMQ and/or HLT (see [05] for details).

We chose not to limit our definitions to necessary and sufficient conditions but rather enlarge them to potentially useful information on the related clinical findings. For example, one may expect to observe *increased troponin* in most patients presenting *myocardial infarct* but this remains empirical knowledge. Such an association is useful in pharmacovigilance to make inferences from a given sign or investigation to a related clinical finding. We applied this strategy to a large part of the SMQ as we intend to reproduce and enable this approach.

3 Results

3.1 Building OntoADR resource

We first extracted 1,186,506 concepts related to SNOMED-CT from UMLS' metathesaurus (325,480 distinct concepts) and 93,308 concepts related to MedDRA (48,363 distinct concepts). Using this extraction and their associated CUI, we were able to find 108,381 mappings between MedDRA and SNOMED-CT (2.24 mappings per MedDRA terms on average), but 5,722 MedDRA preferred terms were not mapped.

We also extracted the hierarchies between concepts from SNOMED-CT (542,485 hierarchical relations) and MedDRA (83,353 hierarchical relations).

Last, from SNOMED-CT we extracted all semantic relationships (except hierarchical properties) from concepts that mapped with MedDRA. This led to 116,527 semantic relationships.

The extraction of SNOMED-CT definitions allowed us to build the list of properties used to define the MedDRA terms. We transformed the SNOMED-CT concepts into 25 semantic relations (see Table 1), i.e. the Finding Site concept was transformed into the 'hasFindingSite' property. The list consists of the 16 SNOMED-CT attributes used to define Clinical Finding concepts, and nine attributes used to define Procedure concepts.

Table 1: List of Semantic Relations used in OntoADR

OntoADR Findings Semantic Relations	***OntoADR Procedures Semantic Relations***
hasFindingSite	hasFindingMethod
hasAssociatedMorphology	hasFindingInformer
associatedWith	hasProcedureSite
↳ occursAfter	↳ hasDirectProcedureSite
↳ dueTo	↳ hasIndirectProcedureSite
↳ hasCausativeAgent	hasMethod
hasSeverity	hasComponent
hasClinicalCourse	hasSpecimen
hasEpisodicity	hasFocus
Interprets	hasDirectSubstance
hasInterpretation	hasIntent
hasPathologicalProcess	(interprets)
hasDefinitionalManifestation	
hasOccurrence	

The most highly used properties are hasFindingSite and hasAssociatedMorphology, which are particularly useful when reasoning on ADR term semantics, for example to query MedDRA terms expressing the same kinds of disorders but distributed in different branches of the MedDRA hierarchy.

The simple algorithm for the automatic creation of properties from the MedDRA label added approximately 8,200 new properties. After the filtering process, 83,267 semantic relationships remained.

The database representing all OntoADR data is approximately 50MB in size, plus 25MB for cache and index files. We obtained 13,703 MedDRA PT with at least one definition out of 20,559 (67%) after filtering, but 6,856 terms were left undefined.

The curation work for the MedDRA terms took approximately 750 hours (equivalent to 5 months as a full time task) and led to 1,935 validated and fully defined terms. The semi-automatic method delivered 3,482 properties for these terms. The curation experts validated 2,636 properties (76%), proposed 350 (10%) more precise terms (i.e. narrower terms in the SNOMED-CT hierarchy) than the automatic proposal, and removed 496 properties (14%). Moreover, the curators manually added 13,675 properties (+393%). This increase must be put in perspective, as the added properties were often not necessary and sufficient conditions. The main categories of the manually added properties concerned DefinitionalManifestation

(40%) and interpretation (20%), whereas validated properties concerned FindingSite (42%) and Morphologies (23%).

3.2 Comparison of precision and recall between OntoADR versions

We compared the results obtained when trying to replicate existing MedDRA groupings with those using the first version of OntoADR by Declerck et al. in 2012 [05]. We selected only topics for which the semantic match between our topic and the existing SMQ or HLT was rated "++" (perfect or quasi perfect semantic match). This resulted in seven safety topics out of 13.

We replicated the queries for each topic and compared the resulting sets of MedDRA terms using precision and recall measures. The recall and precision for OntoADR V1 correspond to results described in [05], and therefore describe a comparison between reference groupings in MedDRA 13.0 and the terms selected by the algorithm. The recall and precision for OntoADR V2 were measured using a comparison with reference groupings in MedDRA 18.0. The comparison between the two approaches is not straightforward as MedDRA 18.0 includes more terms than MedDRA 13.0. Between our first version of OntoADR based on MedDRA 13.0 (which included 18,786 PTs) and the version described in this article based on MedDRA 18.0 (which includes 21,345 PTs), the term count has increased by 14% (+2,559 PTs). The results are presented in Table 2. The column "Variation of PT" indicates the difference in the number of PTs in the reference grouping between MedDRA 13.0 and MedDRA 18.0 for each topic.

Table 2: Comparison of Precision and Recall between two versions of OntoADR for seven safety topics

Safety Topic vs. Reference Grouping	Variation of PT	OntoADR V1		OntoADR V2	
		Recall	Precision	Recall	Precision
Bullous eruptions vs. SMQ Bullous conditions	+3 (+9.1%)	71.9%	52.3%	91.7%	64.7%
Acute renal failure vs. SMQ Acute renal failure	+6 (+15.8%)	5.3%	66.7%	13.6%	60.0%
Aplastic anaemia/pancytopenia vs. HLT Marrow depression	+1 (+4.4%)	66.7%	100.0%	72.7%	88.9%
Neutropenia vs. HLT Neutropenias	+1 (+7.1%)	100.0%	59.1%	100.0%	45.5%
Confusional state vs. HLT Confusion and disorientation	+2 (+66.7%)	100.0%	42.9%	100.0%	25.0%
Thrombocytopenia vs. HLT Thrombocytopenias	+2 (+16.7%)	100.0%	36.7%	100.0%	43.8%
Peripheral neuropathy vs. SMQ Peripheral neuropathy	+9 (+14.8%)	48.3%	24.4%	72.9%	47.7%
	AVERAGE	**70.3%**	**54.6%**	**78.7%**	**53.7%**

We observed a general improvement in recall (+0.084) and a slight decrease in precision (-0.009).

3.3 Use of OntoADR in practice

We tested OntoADR in practice. Several papers have been published about OntoADR in pharmacovigilance applications for term retrieval and signal detection, and several others are currently being considered for publication.

Our first publications were focused on applications for term retrieval. An initial evaluation was performed on 13 medical conditions [05]. Our objective was to replicate the content of already existing groupings (SMQs or HLTs). Selection of these 13 medical conditions was motivated by Trifiro's list of 23 first priority pharmacovigilance safety topics. Originally used for a proof of concept, they serve today as a benchmark to assess the improvement between OntoADR versions (see Table 2).

We have individually dissected several safety topics in greater detail for a better analysis, explaining our methodology and modeling choices. While this article focusses on duplicating the content of SMQs, we have conducted other experiments for which the objective was to match groupings of terms manually selected by a domain expert. For example, recall reached 100% and precision 78.6% for our custom grouping "bullous eruptions" [20]. We identified six additional terms that were absent from the reference SMQ such as 'Oropharyngeal blistering' or 'Tongue Blistering' [20]. We searched for MedDRA PTs related to "Cardiac valve fibrosis", a

safety topic that was investigated after this ADR was observed for Benfluorex in France, and obtained similar results.

These good results for ADR retrieval led us to expand our research in pharmacovigilance databases, especially for signal detection. We used Bayesian statistical methods, such as Empirical Bayes Geometric Mean (EBGM), and indices, such as EB05 (estimated lower 95% "confidence limit" for the EBGM), to detect significant disproportionalities between the number of expected reports and the number of reports observed in pharmacovigilance databases.

With the safety topic "upper gastrointestinal bleedings" [21], we constructed two groupings: (1) we targeted "Hemorrhage" in the "Upper digestive tract structure" and (2) complemented this by taking into account additional MedDRA terms describing the clinical manifestations "Melena" or "Hematemesis". We compared terms in our groupings with our gold standard achieving a recall of 71.0% and a precision of 81.4% for grouping 1; and a recall of 96.7% and a precision of 77.0% for grouping 2. We compared values of the EB05 for 50 randomly selected active ingredients in the public version of the FDA pharmacovigilance database. We observed a coefficient of correlation of R^2 = 0.87 between grouping 1 and the SMQ; and R^2 = 0.99 between grouping 2 and the SMQ, showing good applicability of our custom groupings to signal detection relative to MedDRA groupings.

We obtained similar signal detection results for a second application with "anaphylactic shocks" [22], reaching a precision and recall of 100% for our grouping vs. the corresponding HLT, and 71% recall and 71% precision versus the closest SMQ. We identified two additional terms: 'Anaphylactoid syndrome of pregnancy' and 'First use syndrome' [22]. An updated version of MedDRA has recently introduced this second term into the SMQ.

4 Discussion

4.1 Results analysis

We described (Introduction) our first approach to building OntoADR (V1) and then (in the Methods section) our new approach (V2) that presents several advantages over the first one. Table 3 presents the principal differences between the two approaches.

Table 3: Comparison of V1 and V2 approaches

V1 (Before)	***V2 (After)***
1-to-1 mapping	1-to-n mapping
↳ Manual selection when 1-to-n proposal	↳ Automatic process
↳ Potential loss of definitions	↳ Definitions improved
MedDRA hierarchy preserved	Only MedDRA PTs are defined
↳ Inheritance problems	↳ No inconsistencies
↳ Need for manual corrections	↳ Automatic Inferences
OWL format	Database representation
↳ Complex reasoning	↳ Adapted reasoning
↳ Queries intractable	↳ Negation and Union queries available

The recall increased by 0.084 and precision decreased by 0.009 between V1 and V2. The new version significantly improves data retrieval for the following reasons:

- As explained above, OntoADR V1 was compared to MedDRA 13.0 whereas OntoADR V2 was compared to MedDRA 18.0. Between the two versions, the term count in MedDRA increased by 14%. When measuring recall, its value should mechanically decrease (because when computing the ratio, the denominator increases). Obtaining improved recall is proof of the enhancement of OntoADR.
- The properties of parents are no longer inherited by their children because MedDRA hierarchy is ignored in OntoADR V2. This has decreased the recall when such properties were relevant.
- In V2, we observed a substantial number of terms compared to V1 that, in our opinion, could be included in reference groupings. E.g. “Postoperative renal failure” could be in the Acute Renal Failure SMQ, “Granulocyte count decreased” in HLT Neutropenias, “Confusional arousal” in HLT Confusion and disorientation, “Pancytopenia” in HLT Thrombocytopenias or even “Radiculopathy” in SMQ Peripheral neuropathy. Therefore, the precision measure was negatively affected.

SMQs and HLTs are considered to be reference groupings in MedDRA. It is therefore normal to select them as gold standards, which explains our choice to use them. However, their heterogeneity and the multiple perspectives in their content raised many issues. For example, the HLT Thrombocytopenias (14 PTs) and SMQ Thrombocytopenia (12 PTs) have only three terms in common. Our automatic grouping presents seven terms from the SMQ and the 14 terms from the HLT. We also observed low recall for “acute renal failure”. Indeed,

the corresponding SMQ do not include actual "acute" or "failures" terms, thus our query targeting acute failures for the kidney resulted in low recall. We generally observed better results when making comparisons with custom groupings designed by experts on demand with precise inclusion and exclusion criteria (see above examples in results section).

We expected to obtain improved results with new query features such as negation and disjunction, but their added value does not appear in the results. Indeed, there was a bias in the work reported by [05] because the groupings had not been made entirely automatically. Declerck et al. sometimes manually removed some terms from the list as negation queries were not possible in V1. The new version allows negation queries and the results are fully automated (not modified by any means).

We found that response times were equivalent using our representation in a database or reasoning in OWL. Moreover, the overall calculations were simplified with our approach, which allowed the calculation of negations in a finite time.

4.2 Limits

Our automated approach for defining MedDRA terms enabled us to define 67% of PTs which is encouraging. The terms that benefited from proposed definitions in this automated approach had to be manually evaluated. The curation work performed on our sample showed an error rate of 14%, which is reasonable, but shows that our approach could be improved.

Our automatic processes have mostly allowed us to define properties such as 'hasFindingSite' or 'hasAssociatedMorphology' that are essential for semantic queries (necessary and sufficient conditions), but were less efficient for properties such as hasDefinitionalManifestation that rely on empirical knowledge. After curation, we observed that about 60% of 'hasFindingSite' or 'hasAssociatedMorphology' properties come from automatic methods. This ratio dropped to 25% when all properties were considered. This figure must be put in perspective because the majority of added properties are not necessary and sufficient conditions and are relatively optional for the description of the given medical conditions.

We encountered unexpected difficulty when defining terms. There may be several ways to represent comparable information due to the high potential compositionality of the

SMOMED-CT terms. For example, "hasAssociatedManifestation Peripheral Demyelinating Neuropathy" is equivalent to "hasFindingSite Peripheral Nerve Structure" and "hasMorphology Demyelination". Similarily, "hasMorphology Acute Inflammation" is equivalent to "hasMorphology Inflammation" and "hasClinicalCourse Acute". The way we define these terms can substantially affect the results for users. Our strategy did not initially take into account the choice of either the most granular concept for filling the relations (e.g. favoring Peripheral Demyelinating Neuropathy) or two distinct concepts (Demyelination and Peripheral Nerve Structure). This introduced a bias, as some relations were not retrieved due to this inconsistency in definitions. We are now considering editorial rules to describe a policy for applying compositionality to SNOMED-CT that would take into account any user request. Following Cornet's recommendation describing this same problem, we are also investigating algorithmic rules to transform a given form of expression to another [23].

4.3 Perspectives

Currently available tools dedicated to MedDRA term selection, such as the MedDRA Browser (proposed by the MSSO [24]), only implement string search (searching keywords in the label of terms) and hierarchical browsing. A complementary approach that uses a formal semantic representation of MedDRA could substantially decrease the time needed for term selection and improve the precision of the terms used to describe ADRs. The primary added value of formalizing MedDRA is in enabling automatic term selection based on semantic criteria: for example, to select MedDRA terms with the same FindingSite (e.g. coronary artery), and/or Morphology (e.g. stenosis) independently of MedDRA organization and MedDRA term labels.

We are developing a graphical user interface that allows the user to query the OntoADR resource using a form-based interface. We are conducting an ergonomic evaluation for which the preliminary results are encouraging. This interface will provide users with an improved tool that corresponds to their requirements that is efficient and easy to use.

To be adopted by pharmacovigilance staff, the proposed approach should be implemented in a commercial tool. We plan to contact software editors and help them accompany users for the adoption of these innovative techniques. Improving MedDRA

terminology searches will result in better groupings for signal detection, so we also plan to contact signal detection software providers.

A comparison of formal definitions available in OAE with our formal definitions, via cross referencing with MedDRA, would help to improve our resource by evaluating discrepancies between both approaches. However, this is not straightforward because a) anatomy concepts in OAE are defined in UBERON that integrates parts of the FMA (chosen to apply to both human and veterinary medicine) that we would need to first map to SNOMED-CT; 2) pathological bodily processes should be mapped on findings or morphologies of SNOMED-CT; and 3) OAE and MedDRA probably do not have the same granularity (i.e. OAE is often broader than MedDRA). Moreover, there is no information other than anatomical entities and pathological processes in OAE, so we would not be able to compare the other relationships.

Ingenerf has suggested that formal systems should complement current terminological systems rather than replace them [25]. We followed this advice and expect that our approach could be generalized to other terminologies that do not benefit from a formal system. Recently, Lee proposed that using post-coordinated expressions may be a convenient way to retrieve ICD codes but observed that such an approach is still untested [26]. We favored this approach to improve searches of MedDRA and agree that this methodology would be useful for other terminologies such as ICD.

We recommended submitting any developments based on MedDRA, such as OntoADR, for the sustainability and the legality of our resource, as MedDRA is being developed and maintained by the MSSO. The MSSO has been contacted and is aware of OntoADR. Closely coordinated development would facilitate the evolution of OntoADR which must keep pace with the evolution of MedDRA (two releases per year). The adaptation of the resource to introduce new terms and potentially their automatic definition through external sources is already supported. For example, modification of SNOMED-CT (which is also evolving twice a year) that occurs is transmitted in a MedDRA term definition is marked as non-curated. Likewise, a new SNOMED-CT to MedDRA mapping can immediately be imported into OntoADR.

Overall our methods proved to be very efficient with half of MedDRA preferred terms but were insufficient to achieve coverage of all MedDRA terms. In order to rely more on automation, current processes should be improved using several methods: 1) Parsing textual

material by acquiring formal definitions from textual definitions [27] or directly from the term label using morpho-semantic analysis, e.g. 'ectomy' stands for 'Surgical excision' or 'gastr' stands for 'stomach'. Such approach is limited to terms containing "compound forms" that have a medical meaning [28]. 2) Using ontology design patterns (ODP) and tools like Ontorat [29] or TermGenie [30] that are in a way similar to Ci4SeR [18] and rely on expert curation. 3) Using other sources or improving existing ones: Diallo et al [31], Fung et al. [32] or Bodenreider et al. [33] developed methods to improve mappings already available in UMLS. In a similar way we performed preliminary experiments to extract knowledge from NCI Thesaurus [34] to build more MedDRA definitions in the field of cancer related adverse reactions. Or 4) Auditing semantic definitions by comparing definitions associated to terms that present lexical similarities [35]. But this presents an intrinsic limit: terms to compare should consist of at least three words which constraints this method mainly to MedDRA procedures.

5 Conclusion

Here, we have presented a new methodology for building a semantic resource consisting of formal definitions of ADRs (OntoADR). Searches in MedDRA were previously limited to hierarchical browsing and syntactical searches. Although ontologies are available in the medical domain to improve data retrieval, the field of pharmacovigilance does not yet benefit from a fully operational ontology to formally represent the MedDRA terms and enable reasoning. OntoADR allows users to perform a new kind of search based on semantic definitions, which enables a complementary approach for data retrieval within ADR terms. We previously demonstrated the relevance and advantages of such an approach. Our new methodology for building OntoADR further improves data retrieval with respect to specificity, sensibility, flexibility, and efficiency relative to our previous approach.

In the improved approach, we built a semantic resource for ADRs, based on MedDRA reference terminology, using a database that allows better curation. This adds improved search capabilities relative to our previous approach and even overcomes EL++ limitations, such as the inability to perform negation and conjunction, bringing it closer to user need. Further curation efforts are still necessary to improve our resource.

Data retrieval in terminologies such as MedDRA can be improved using resources such as OntoADR and this can pave the way to new domains that can develop based on such enhanced searches. This result is promising, because MedDRA term groupings for pharmacovigilance (mainly SMQs) are currently performed manually, and even partial automation of the process, would lead to substantial time savings.

Acknowledgments

The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi-protect.eu) which is a public-private partnership coordinated by the European Medicines Agency. The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement n° 115004, resources which is composed of financial contributions from the European Union's Seventh Framework Program (FP7/2007-2013) and in-kind contributions of EFPIA companies.

We acknowledge Eric Sadou, Adrien Fanet, and Anne Jamet who contributed to the development of OntoADR.

The views expressed are exclusively those of the authors.

References

[01] Brown EG. Methods and pitfalls in searching drug safety databases utilising the Medical Dictionary for Regulatory Activities (MedDRA). Drug Saf. 2003;26(3):145-58.

[02] Cimino JJ. In defense of the Desiderata. J Biomed Inform. 2006 Jun;39(3):299-306.

[03] Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). Drug Safety, 20(2), 109-117.

[04] Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. Methods of information in medicine, 37(4-5), 394.

[05] Declerck, G., Bousquet, C., & Jaulent, M. C. (2012). Automatic generation of MedDRA terms groupings using an ontology. In MIE (pp. 73-77).

[06] Bousquet, C., Sadou, É., Souvignet, J., Jaulent, M. C., & Declerck, G. (2014). Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms. Journal of biomedical informatics, 49, 282-291.

[07] Baader, F., Brandt, S., & Lutz, C. (2005, July). Pushing the EL envelope. In IJCAI (Vol. 5, pp. 364-369).

[08] Baader, F., Brandt, S., & Lutz, C. (2008). Pushing the EL envelope further.

[09] Rector, A. L., & Brandt, S. (2008). Why do it the hard way? The case for an expressive description logic for SNOMED. Journal of the American Medical Informatics Association, 15(6), 744-751.

[10] Blondé, W., Mironov, V., Venkatesan, A., Antezana, E., De Baets, B., & Kuiper, M. (2011). Reasoning with bio-ontologies: using relational closure rules to enable practical querying. Bioinformatics, 27(11), 1562-1568.

[11] Schulz, S., Suntisrivaraporn, B., Baader, F., & Boeker, M. (2009). SNOMED reaching its adolescence: Ontologists' and logicians' health check. International journal of medical informatics, 78, S86-S94.

[12] Taboada, M., Martínez, D., Pilo, B., Jiménez-Escrig, A., Robinson, P. N., & Sobrido, M. J. (2012). Querying phenotype-genotype relationships on patient datasets using semantic web technology: the example of cerebrotendinous xanthomatosis. BMC medical informatics and decision making, 12(1), 78.

[13] Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. Methods of information in medicine, 32(4), 281-291.

[14] Nadkarni, P. M., & Darer, J. D. (2010). Determining correspondences between high-frequency MedDRA concepts and SNOMED: a case study. BMC medical informatics and decision making, 10(1), 66.

[15] Date, C. J., & Darwen, H. (1993). A guide to the SQL Standard: a user's guide to the standard relational language SQL (Vol. 55822). Addison-Wesley Longman.

[16] Maier, D. (1983). The theory of relational databases (Vol. 11). Rockville: Computer science press.

[17] Handy, J. (1998). The cache memory book. Morgan Kaufmann.

[18] Souvignet, J., Asfari, H., Declerck, G., Lardon, J., Trombert-Paviot, B., Jaulent, M. C., & Bousquet, C. (2014). Ci4SeR–Curation Interface for Semantic Resources–Evaluation with Adverse Drug Reactions. EHealth-For Continuity of Care: Proceedings of MIE2014, 205, 116.

[19] Trifirò, G., Pariente, A., Coloma, P. M., Kors, J. A., Polimeni, G., Miremont-Salamé, G., ... & Fourrier-Reglat, A. (2009). Data mining on electronic health record databases for signal

detection in pharmacovigilance: which events to monitor?. Pharmacoepidemiology and drug safety, 18(12), 1176-1184.

[20] Asfari H., Souvignet J., Guy C. and Bousquet C. - Knowledge-based method for automated generation of new MedDRA Grouping for bullous conditions, P2T 2014

[21] Souvignet J., Declerck G., Jaulent M-C., Bousquet C. - Evaluation of Automated Term Groupings for Detecting Upper Gastrointestinal Bleeding Signals for Drugs, Drug Safety 2012; 35 (12): 1195-6

[22] Souvignet J., Declerck G., Trombert-Paviot B., Rodrigues J-M., Jaulent M-C., Bousquet C. - Evaluation of automated term groupings for detecting anaphylactic shock signals for drugs, AMIA Annu Symp Proc. 2012; 882-90

[23] Cornet, R., Nyström, M., & Karlsson, D. (2013, August). User-directed coordination in SNOMED CT. In MedInfo (pp. 72-76).

[24] http://www.meddra.org/browsers [date accessed: 11/01/2016]

[25] Ingenerf, J., & Giere, W. (1998). Concept-oriented standardization and statistics-oriented classification: continuing the classification versus nomenclature controversy. Methods of information in medicine, 37(4-5), 527-539.

[26] Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform. 2013;46(1):87-96..

[27] Petrova A, Ma Y, Tsatsaronis G, Kissa M, Distel F, Baader F, Schroeder M. Formalizing biomedical concepts from textual definitions. J Biomed Semantics. 2015;6:22.

[28] Deléger L, Namer F, Zweigenbaum P. Morphosemantic parsing of medical compound words: transferring a French analyzer to English. Int J Med Inform 2009;78Suppl 1:S48-55.

[29] Xiang Z, Zheng J, Lin Y, He Y. Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. J Biomed Semantics. 2015;6:4.

[30] Dietze H, Berardini TZ, Foulger RE, Hill DP, Lomax J, Osumi-Sutherland D, Roncaglia P, Mungall CJ. TermGenie - a web-application for pattern-based ontology class generation. J Biomed Semantics. 2014;5:48.

[31] Diallo G. An effective method of large scale ontology matching. J Biomed Semantics. 2014;5(1):44.]

[32] Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. Stud Health Technol Inform. 2007;129(Pt 1):605-9.

[33] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proceedings / AMIA Annual Symposium. 1998:815–9.

[34] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30-43.

[35] Agrawal A, Elhanan G. Contrasting lexical similarity and formal definitions in SNOMED CT: consistency and implications. J Biomed Inform. 2014;47:192-8.

Conflict of interest

No conflict to declare

Highlights

- OntoADR a semantic resource describing MedDRA terms is proposed.
- A formal definition is proposed for 67% of MedDRA preferred terms.
- OntoADR enables a new kind of criteria based data retrieval.
- Sensitivity and specificity are improved compared to previous approaches.

Graphical abstract