Ontological and Non-Ontological Resources for Associating Medical Dictionary for Regulatory Activities Terms to SNOMED Clinical Terms With Semantic Properties

Cédric Bousquet[1,2]*, Julien Souvignet[1,2], Éric Sadou[1], Marie-Christine Jaulent[1] and Gunnar Declerck[3]

[1] Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Sorbonne Université, Inserm, Université Paris 13, Paris, France, [2] Unit of Public Health and Medical Informatics, University of Saint Etienne, Saint Etienne, France, [3] EA 2223 Costech (Connaissance, Organisation et Systèmes Techniques), Centre de Recherche, Sorbonne Universités, Université de technologie de Compiègne, Compiègne, France

Background: Formal definitions allow selecting terms (e.g., identifying all terms related to "Infectious disease" using the query "has causative agent organism") and terminological reasoning (e.g., "hepatitis B" is a "hepatitis" and is an "infectious disease"). However, the standard international terminology Medical Dictionary for Regulatory Activities (MedDRA) used for coding adverse drug reactions in pharmacovigilance databases does not beneficiate from such formal definitions. Our objective was to evaluate the potential of reuse of ontological and non-ontological resources for generating such definitions for MedDRA.

Methods: We developed several methods that collectively allow a semiautomatic semantic enrichment of MedDRA: 1) using MedDRA-to-SNOMED Clinical Terms (SNOMED CT) mappings (available in the Unified Medical Language System metathesaurus or other mapping resources, e.g., the MedDRA preferred term "hepatitis B" is associated to the SNOMED CT concept "type B viral hepatitis") to extract term definitions (e.g., "hepatitis B" is associated with the following properties: has finding site liver structure, has associated morphology inflammation morphology, and has causative agent hepatitis B virus); 2) using MedDRA labels and lexical/syntactic methods for automatic decomposition of complex MedDRA terms (e.g., the MedDRA systems organ class "blood and lymphatic system disorders" is decomposed in blood system disorders and lymphatic system disorders) or automatic suggestions of properties (e.g., the string "cyclic" in preferred term "cyclic neutropenia" leads to the property has clinical course cyclic).

Results: The Unified Medical Language System metathesaurus was the main ontological resource reusable for generating formal definitions for MedDRA terms. The non-ontological

resources (another mapping resource provided by Nadkarni and Darer in 2010 and MedDRA labels) allowed defining few additional preferred terms. While the Ci4SeR tool helped the curator to define 1,935 terms by suggesting potential supplemental relations based on the parents' and siblings' semantic definition, defining manually all MedDRA terms remains expensive in time.

Discussion: Several ontological and non-ontological resources are available for associating MedDRA terms to SNOMED CT concepts with semantic properties, but providing manual definitions is still necessary. The ontology of adverse events is a possible alternative but does not cover all MedDRA terms either. Perspectives are to implement more efficient techniques to find more logical relations between SNOMED CT and MedDRA in an automated way.

Keywords: adverse drug reaction, Medical Dictionary for Regulatory Activities, SNOMED Clinical Terms, ontology, clinical terminology, pharmacovigilance

INTRODUCTION

Formal representation of semantics as provided by computational ontologies and associated semantic Web techniques have been extensively used in medical data integration systems in the last decade (Sheth et al., 2005), and they now tend to be acknowledged as a powerful means to improve the quality of the processing chain of medical data, process automatic extraction of information and knowledge from large databases or ensure semantic interoperability between disparate data processing systems (Park and Hardiker, 2009; Schriml et al., 2012; Schulz and Jansen, 2013).

In the medical domain, classic terminologies are gradually giving way to clinical terminologies, in which terms are defined using knowledge representation languages (Rossi Mori et al., 1998). An example is SNOMED Clinical Terms (SNOMED CT), a general clinical terminology whose objective is to represent all possible terms required for coding the patient record and other applications for representation of biomedical information (Khorrami et al., 2018). SNOMED CT presents several advantages compared with classic terminologies, especially the ability to apply techniques of semantic reasoning in order to build new groups of terms, whereas classic terminologies are limited to default groupings (generally made manually by experts) that are already specified as part of the terminology (Bousquet et al., 2005).

Medical Dictionary for Regulatory Activities (MedDRA) is a classic terminology used by regulatory authorities and pharmaceutical companies for coding adverse drug reactions (ADR) in pharmacovigilance databases (Brown et al., 1999). MedDRA terms are not formally defined and search is therefore limited to existing categories (Bousquet et al., 2005). It is frequently difficult to identify the exact MedDRA category that represents a given medical condition under investigation in a sufficiently specific and exhaustive way, for example, during a pharmacovigilance database search (Brown, 2003).

Since several years, we have performed studies that showed that a knowledge-based approach is efficient for building new groups of ADR terms with World Health Organization Adverse Reaction Terminology (WHO ART) (Alecu et al., 2006) (Iavindrasana et al., 2006) and with MedDRA (Henegar et al., 2006; Declerck et al., 2012; Asfari et al., 2016; Souvignet et al., 2016a) in an automated way. This means that starting from a resource containing formal definitions of ADR terms, it is possible to make queries that correspond to a case definition in order to retrieve the related set of terms. This strategy was applied in Pharmacovigilance Adverse Reaction Terminology Server (Alecu et al., 2007) where building a knowledge base for all WHO ART terms was a challenge. Indeed, all definitions were to be set manually, and we therefore focused on automated ways to enrich WHO ART (Iavindrasana et al., 2006). We found that mapping of WHO ART with SNOMED CT by means of the Unified Medical Language System (UMLS) metathesaurus proved to be a very efficient method to build formal definitions of WHO ART terms in an automated way (Alecu et al., 2008).

Difficulties we encountered for enriching WHO ART now appear at a larger scale in MedDRA due to a growing number of terms and a more complex organization of MedDRA. Indeed, only about 50% of MedDRA terms [excluding lowest level term (LLT)] were associated with a SNOMED CT concept in UMLS (Bodenreider, 2009). Therefore, the mapping method we applied to WHO ART was a fair starting point but proved to be insufficient for obtaining an exhaustive enrichment of MedDRA.

Our objective was to evaluate the potential of reuse of ontological and non-ontological resources for defining and/or enriching definitions of MedDRA terms. We present in this article

---

Abbreviations: ADR, Adverse drug reactions; AERS, Adverse Events Reporting System; Ci4SeR, Curation Interface for Semantic Resources; HLGT, Higher Level Group Term; HLT, High Level Term; ICD-10, International classification of diseases, 10th edition; LALR, Lexically assign logically refine; LLT, Lowest Level Term; LOINC, Logical Observation Identifiers Names & Codes; MedDRA, Medical dictionary for drug regulatory activities; NEC, Not elsewhere classified; NOS, Not otherwise specified; OAE, Ontology of Adverse Events; OWL, Web Ontology Language; PT, Preferred Term; SMQ, Standardized MedDRA Queries; SNOMED CT, SNOMED Clinical Terms; SOC, System Organ Class; UMLS, Unified Medical Language System; WHO ART, World Health Organization-Adverse Reaction Terminology.

several complementary methods that may benefit different levels of automation and could also be reused in order to semantically enrich other terminologies. These include methods such as i) extracting SNOMED CT definitions based on MedDRA-to-SNOMED CT mappings available in the UMLS metathesaurus or other mapping resources and ii) developing lexical and syntactic methods using MedDRA term label information. The ability to reuse the selected ontological and non-ontological resources was measured by comparing the number of MedDRA terms associated with a formal definition after processing of these resources with the number of MedDRA terms to define. Additionally, we manually curated some term definitions using expert knowledge that allowed us to evaluate the time necessary and validated a sample of the formal definitions provided by the previous methods. We stored formal definitions of MedDRA terms in a semantic resource named OntoADR (Bousquet et al., 2014).

The organization of OntoADR and results of semantic queries on OntoADR have been already published (Bousquet et al., 2014; Souvignet et al., 2016b). This article presents methods we implemented for reusing ontological and non-ontological resources to enable the formalization of the semantics and results obtained with each of these methods, how they automate the development of formal representations of MedDRA terms, the limits related to these methods, and additional developments that would be required for a more complete semantic enrichment of MedDRA.

BACKGROUND

Hierarchical Organization of Medical Dictionary for Regulatory Activities

The MedDRA hierarchy consists of five levels (from broad to narrow), among which four are depicted in **Figure 1**: System Organ Class (SOC), e.g., hepatobiliary disorders; higher level group term (HLGT), e.g., hepatic and hepatobiliary disorders; high level term (HLT), e.g., hepatic viral infections; preferred term (PT), e.g., hepatitis B; and LLT not shown on the figure. The PT level is preferred for data analysis and retrieval. MedDRA was defined as multi-axial because one PT may be present in one primary SOC and also in several secondary SOC. However, one PT may exist only within one single HLT within a SOC. As HLT within a SOC constitutes disjoint classes, it is seldom reliable to consider only one HLT or higher level category when searching for MedDRA terms related to a pharmacovigilance safety topic (Bousquet et al., 2005; Asfari et al., 2016).

Moreover, it was recognized that HLT are not always sufficient to represent clinical conditions involving several organs (e.g., anaphylactic shock involving the kidney, liver, cardiovascular, and respiratory systems) because they only group together terms belonging to the same SOC. When searching for signals associated with a drug, MedDRA terms representing the suspected ADR must thus be identified prior to the running of signal detection algorithms Souvignet et al. (2012). For instance, if one suspects a given drug to cause acute renal failure, using the MedDRA term "renal failure acute" is generally not sufficient for the algorithms to extract a signal because the acute renal failure condition can be coded with several related MedDRA terms by health professionals (e.g., "renal impairment," "blood creatinine abnormal," or "dialysis"). Identifying clinically related terms in MedDRA is not an easy task, as those terms might exist in different locations of the MedDRA hierarchy.

Since several years, the Maintenance and Support Services Organization, which is responsible for MedDRA maintenance and diffusion, builds standardized MedDRA queries (SMQ) to address these issues (Mozzicato, 2007). SMQs consist of sets of PT from different branches of MedDRA that allow describing a particular medical condition and are intended to aid in case identification. SMQs are a way to describe safety topics relevant
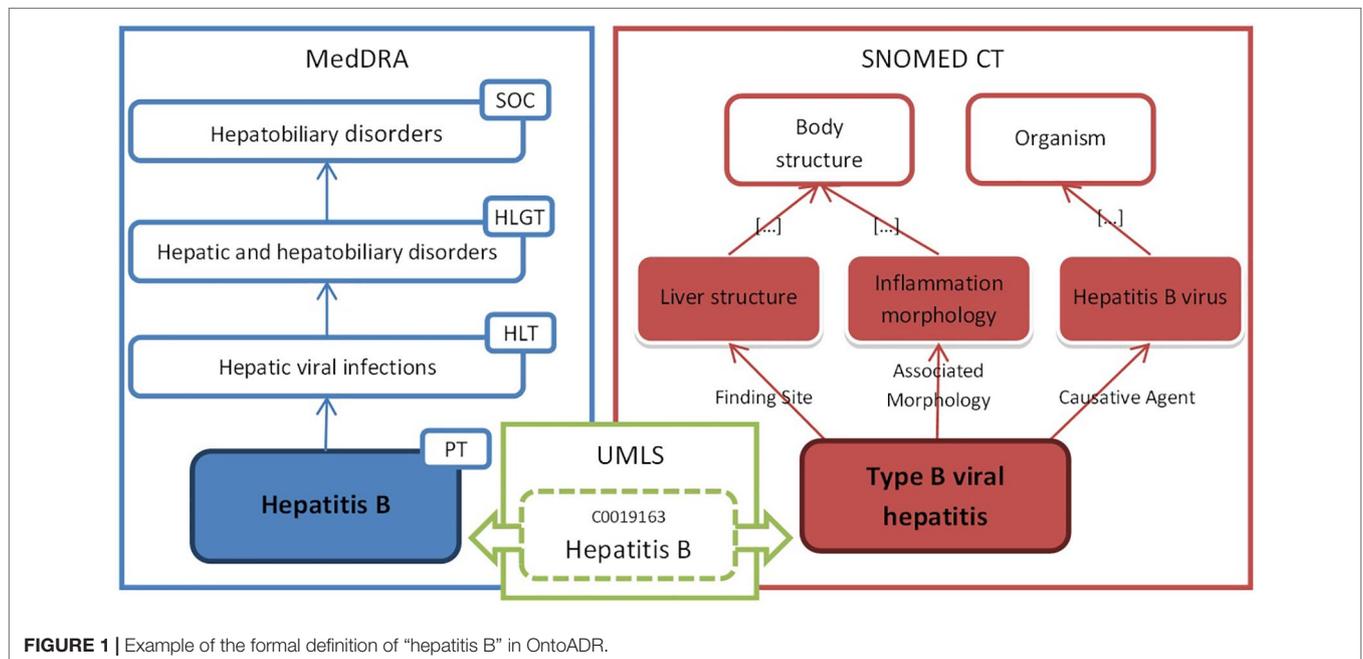

FIGURE 1 | Example of the formal definition of "hepatitis B" in OntoADR.

for pharmacovigilance that are not covered by the HLT and HLGT present in the MedDRA hierarchy. However, the achievement of SMQ raises important difficulties.

SMQs are currently developed manually by experts from the Maintenance and Support Services Organization, which is time consuming. Furthermore, once defined, they should not be modified or customized, and the existing SMQs do not cover all issues possible with drugs. Because experts have (even slightly) different understandings of the medical condition targeted by an SMQ, the kind of terms and the rationale for their selection may differ from an SMQ to another. This means that from one group of experts to another, for the same safety topic, the list of MedDRA terms selected could be different. Because SMQs are manually implemented, they could also miss important MedDRA terms. For those different reasons, the development of methods for automated selection of MedDRA terms on the basis of semantic information is desirable. Indeed, an automation of the process of PT selection in SMQ, even partial, could increase the quality and reproducibility of the SMQ and allow an important saving of time.

Difficulties for Searching Terms in Medical Dictionary for Regulatory Activities

The performance of pharmacovigilance systems based on spontaneous reporting is dependent on the information systems in which case reports are stored. In particular, these systems are subordinated to the ability of users to retrieve and exploit case reports in order to 1) reinforce existing knowledge on drug

safety, 2) make assumptions about the existence of a causal relationship between a drug and an adverse event, and 3) evaluate the available information to implement regulatory measures to secure drug therapies. The search for pharmacovigilance case reports is difficult because it is necessary to identify the medical terms indicating the safety topic that one wishes to evaluate. In general, a term is not sufficient to designate this safety topic, and it is preferable to look for all case reports in relation to a set of terms (Hauben et al., 2006; Hansen et al., 2007). According to the MedDRA® Data Retrieval and Presentation: Points To Consider (ICH Working Group, 2018), "clinically related PTs might be overlooked or not recognized as belonging together because they might be in different groupings within a single SOC or they may be located in more than one SOC."

Figure 2 shows the problem of finding terms in MedDRA. Terms associated with a green tick are related to valvulopathy, while terms marked with a red cross do not correspond to valvulopathy. It is observed that the search terms are located in different branches of the terminology, which requires the pharmacovigilant specialist more time and effort to carry out his query. In addition, several HLT or HLGT must be combined to arrive at the final result, and irrelevant terms are present in these groups, which means that a search based on HLT or HLGT groupings will be associated with a large number of irrelevant PT. Another method for searching MedDRA terms is the textual query, but the terminology seems complex and does not reveal discriminating strings in the search for valvulopathies. For example "stenos * aort *" gives as results "stenosis of the aortic valve," "congenital aortic stenosis of the
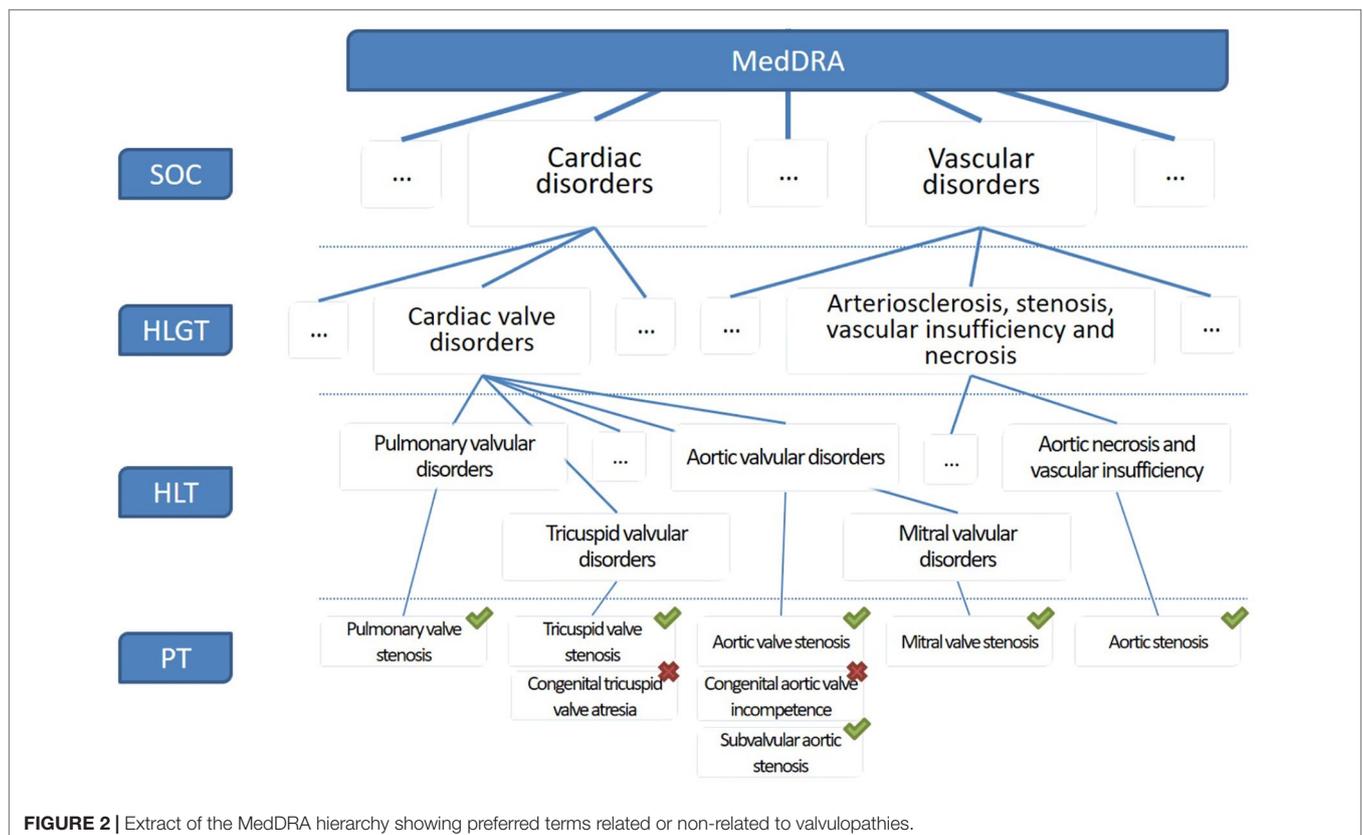

FIGURE 2 | Extract of the MedDRA hierarchy showing preferred terms related or non-related to valvulopathies.

valve," and "stenosis of the mitral valve and insufficiency of the aortic valve," but it is necessary to perform additional text searches corresponding to findings for valvular involvement such as calcification or insufficiency.

Interface, Aggregation, and Reference Terminologies

Schulz et al. (2017) recently evaluated how interface, aggregation, and reference terminologies may interact in the context of the new 11th version of the International Classification of Diseases and its relations with SNOMED CT. Interface terminology was defined by Rosenbloom et al. (2006) as a "systematic collection of health care–related phrases (terms) that supports clinicians' entry of patient-related information into computer programs […]." According to Spackman (1997), the "main purpose of a reference terminology […] is the retrieval and analysis of data." The term "aggregation terminology" was first introduced by Rogers (2005) to designate a classification systems in which its main purpose is to enable "statistical aggregation," further defined by Schulz et al. (2017) as consisting of single hierarchies and disjoint classes.

We consider that such clarification would be useful in our case study, to better explain what we intend to do with MedDRA and SNOMED CT. Within our approach, MedDRA is the aggregation terminology, and SNOMED CT is the reference terminology. As our purpose is to improve retrieval and not coding of MedDRA terms, we did not work on building an interface terminology. Such interface terminology would be desirable to facilitate data entry in pharmacovigilance databases but is outside of our scope. In the following paragraph, we show how a graphical user interface implementing SNOMED CT as a reference terminology could help users' experience in selecting MedDRA terms and potentially improving search in pharmacovigilance databases.

Rationale for Supplementing Medical Dictionary for Regulatory Activities With Formal Definitions

We consider it is possible to overcome the limitations associated with the organization of MedDRA terminology and the difficulty to identify related MedDRA terms, by proposing an alternative method for the grouping of PTs based on their medical meaning rather than their position in the hierarchy. This new method is based on PT modeling in a form that allows logical inferences by a computer. From a technical point of view, the implementation of this method is based on knowledge engineering, a branch of artificial intelligence in which it is possible to describe MedDRA terms using a formal language (Bousquet et al., 2014). In the field of knowledge engineering, we define "ontology" as the set of objects of a domain and relations between these objects.

While McKnight (1999) recognized the need for user-directed composition of controlled health terminologies, and the required improvement of the user interface in the context of data entry, we believe that such user interface is also of great importance to enable composition for data retrieval. In a previous work, we implemented OntoADR query tools, a graphical user interface that relies on OntoADR, and compared the performances of

eight users in selecting MedDRA PTs with the MedDRA web browser in a pilot study on five medical conditions (Souvignet et al., 2019). Although the number of medical conditions was low, we observed a statistically significant improvement by using OntoADR query tools compared with the MedDRA web browser for selecting MedDRA PTs (+27% precision and +34% recall). Similar to Maedche and Staab (2001), we consider that the target application may serve as a measure for validating the implemented ontology and believe that such criteria is more important than criteria relying only on the evaluation of the ontology without taking into account the context where it is used. This pilot study confirmed the validity of our approach and justifies that we continue the implementation of our mappings between MedDRA and SNOMED CT.

In order to address a safety issue, it is necessary to identify case reports in pharmacovigilance databases relative to this issue. A safety issue may concern the causality assessment of drug D in the occurrence of medical condition C. **Figure 3** shows the example of three use cases where one is evaluating the causal role of suspected drug D in three medical conditions: a) upper gastrointestinal hemorrhage, b) medical conditions with symptom of erythema, and c) fungal infectious disorders. A single MedDRA term is usually not sufficient to characterize a given medical condition. OntoADR is intended to support selection of MedDRA terms according to different criteria. **Figure 4** depicts the query performed in OntoADR to retrieve MedDRA terms associated to these medical conditions, e.g., "finding site: upper gastrointestinal tract," and "associated morphology: hemorrhage," and the 10 first MedDRA terms retrieved by this query, e.g., anastomotic ulcer hemorrhage, aorto-esophageal fistula, chronic gastrointestinal bleeding, etc.

Table 1 shows parts of formal definitions for 10 MedDRA terms associated with upper gastrointestinal hemorrhage among 27, in particular, SNOMED CT concepts that are filler of relations "finding site" and "associated morphology." Fillers that are not relevant to upper gastrointestinal hemorrhage are in italic, e.g., *Stenosis* for the MedDRA PT "anastomotic ulcer hemorrhage." In case the filler of the "associated morphology" relation is "hemorrhage," the query is immediately satisfied for this condition, e.g., three MedDRA terms (aorto-esophageal fistula, gastric antral vascular ectasia, and gastric hemorrhage) are defined as having "hemorrhage" as their associated morphology. When "hemorrhage" is not the filler of the "associated morphology" relation, other relevant fillers may be retrieved thanks to the subsumption mechanism that establishes hierarchical relations between a parent concept and its children concept. For example, "hemorrhage" subsumes "acute bleeding ulcer," 'bleeding varices," "chronic hemorrhage," and "hemorrhagic inflammation." "Upper gastro intestinal tract structure" subsumes "duodenal structure," "esophageal structure," "gastrojejunal junction structure," "pyloric antrum structure," and "stomach structure."

Semantic Enrichment

The traditional process of domain ontology construction is based on expert intervention (Bedini and Nguyen, 2007). Although this manual procedure guarantees a fair quality of the generated
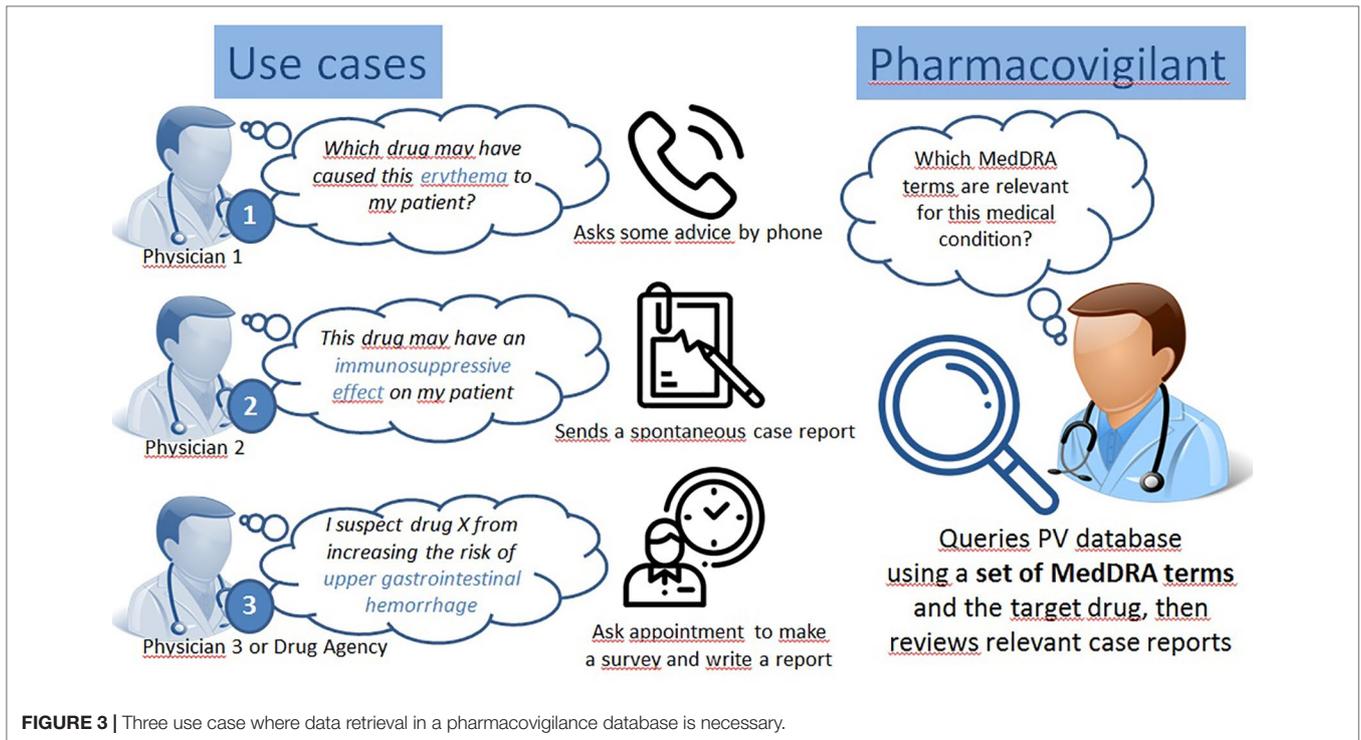
FIGURE 3 | Three use case where data retrieval in a pharmacovigilance database is necessary.
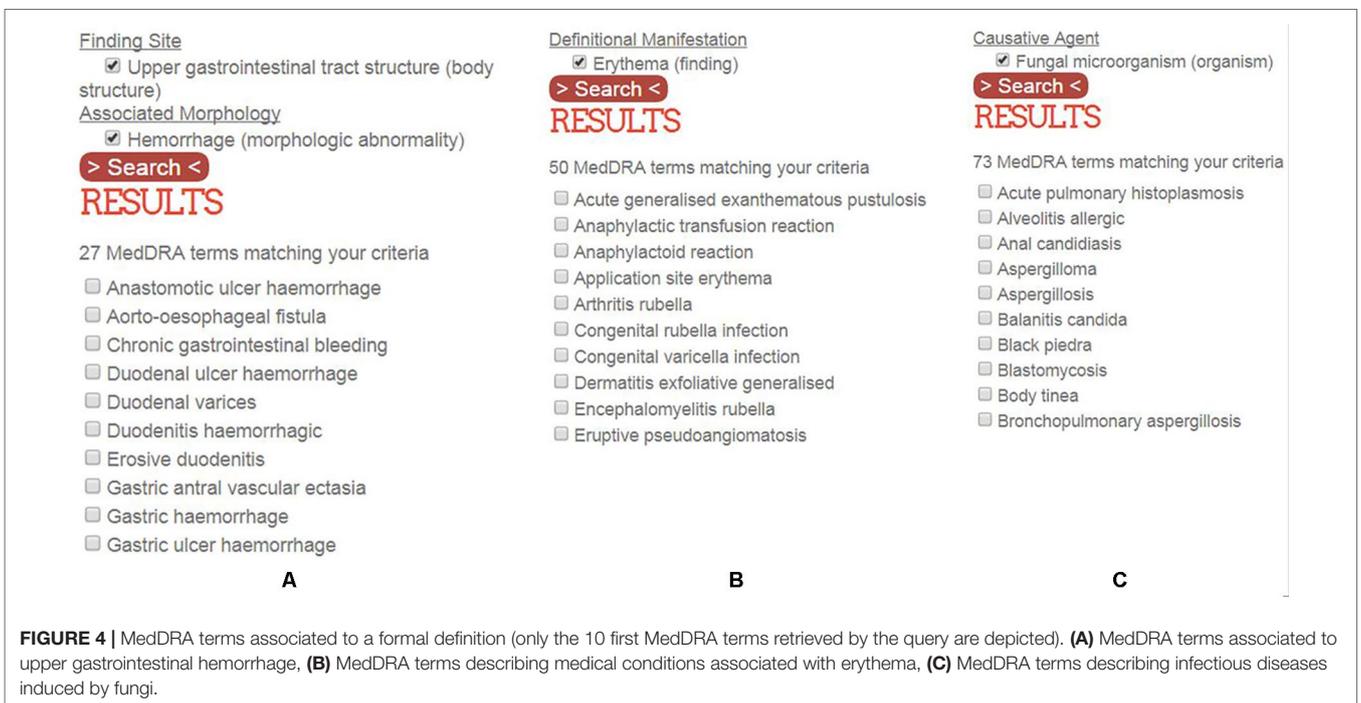

FIGURE 4 | MedDRA terms associated to a formal definition (only the 10 first MedDRA terms retrieved by the query are depicted). **(A)** MedDRA terms associated to upper gastrointestinal hemorrhage, **(B)** MedDRA terms describing medical conditions associated with erythema, **(C)** MedDRA terms describing infectious diseases induced by fungi.

resource, it suffers from several difficulties; among those are the cold start problem (starting from scratch) and the lack of availability of domain experts (Qawasmeh et al., 2018). In fact, the high cost of experts' interventions is the major bottleneck identified early in the state of the art of ontology construction (Cullen and Bryman, 1988; Simperl et al., 2006; Balakrishna

et al., 2010). This bottleneck justifies reusing and linking existing resources, when available, to create new ontologies (Alani, 2006). Reuse is not always possible because ontologies may not exist in the field of interest. For example, Mazo et al. (2017) describe a histological ontology of the human cardiovascular system and report that they "did not find in the State-of-the-Art an ontology of

TABLE 1 | Finding site and associated morphology of 10 MedDRA terms describing upper gastrointestinal hemorrhage among 27.

	hasFindingSite	hasAssociatedMorphology
Anastomotic ulcer haemorrhage	Gastrojejunal junction structure	Acute bleeding ulcer Stenosis
Aorto-esophageal fistula	Aortic structure Esophageal structure	Hemorrhage
Chronic gastrointestinal bleeding	Stomach structure	Chronic hemorrhage
Duodenal ulcer haemorrhage	Duodenal structure	Acute bleeding ulcer
Duodenal varices	Portal vein structure Duodenal structure	Bleeding varices
Duodenitis hemorrhagic	Duodenal structure	Hemorrhagic inflammation
Erosive duodenitis	Duodenal structure	Hemorrhagic inflammation
Gastric antral vascular ectasia	Pyloric antrum structure	Hemorrhage Angiectasia
Gastric hemorrhage	Stomach structure	Hemorrhage
Gastric ulcer hemorrhage	Stomach structure	Acute bleeding ulcer

histology neither a similar organization of hierarchies of histology terms that [they] may be able to reuse." At the opposite, when all the ontologies that are needed are available, it is sufficient to reuse and assemble them, such as in the example of the development of the orthology ontology (Fernández-Breis et al., 2016).

Ontology enrichment is the task of extending an existing ontology with additional concepts and semantic relations and placing them at the correct position in the ontology (Petasis et al., 2011). Automatic ontological construction is often based on learning (Maedche and Staab, 2001; Buitelaar et al., 2005). Such approach can be based on unstructured texts (Asim, 2018; Cimiano et al., 2006; Emani et al., 2015; Costa et al., 2016; Dasgupta et al., 2018), informal ontologies (Astrakhantsev and Turdakov, 2013), or linked data (Gavankar et al., 2012; Tiddi et al., 2012; Riga et al., 2017). A particular case of unstructured data corresponds to the labels of ontology identifiers that may be very dense with information. Such information described as "hidden semantics" by Third (2012) received little attention, at the exception of the gene ontology identifiers (e.g., Quesada-Martínez et al., 2015). SNOMED CT identifiers may also benefit from such approach, but this was limited to taking into account the "acute" and "chronic" qualifiers for evolution of diseases (Rector and Iannone, 2012) and the occurrence of congenital diseases (van Damme et al., 2018). Such "hidden semantics" were also detected by Nadkarni and Darer (2010) to build correspondences between MedDRA and SNOMED CT, but these correspondences were not associated with relations, which makes this work interesting for reuse but requires reengineering to transform the correspondences into semantic relations.

One of the major difficulties of these approaches is the extraction of non-taxonomic relationships (Dahab et al., 2008; Sánchez and Moreno, 2008; Villaverde et al., 2009; Petasis et al., 2011; Serra

et al., 2014). Furthermore, several automatic approaches for ontology reusing and engineering still require domain experts and knowledge engineers (Bobed et al., 2012). Thus, semiautomatic approaches could be a good alternative (Balakrishna et al., 2010).

Semiautomatic approaches employ intelligent methods to significantly reduce, without completely replacing, human efforts (Huang et al., 2014). In such approaches, the role of experts could be limited to validating final automatic learning results (Wächter and Schroeder, 2010) or suggesting improvements at the end of ontology life cycle (Alobaidi, 2018). Expert intervention can be achieved with the help of graphical user interface (Wächter and Schroeder, 2010), spreadsheets (Blfgeh et al., 2017; Judkins et al., 2018), or specified pipelines such as the eXtensible ontology development (He et al., 2018).

NeON Methodology

After comparing several methods for ontology development, we selected the NeON methodology (Suárez-Figueroa et al., 2012) because it was the most appropriate to illustrate the strategy that we followed for designing OntoADR. While other methodologies may also be considered for knowledge engineering and may be more relevant in other contexts, we considered that dimensions of reuse were the most important features when selecting NeON.

While other approaches for ontology engineering provide methodological guidance, the NeON Methodology does not prescribe a rigid workflow. It suggests a variety of pathways based on nine flexible scenarios that address common issues, such as reusing, reengineering, and merging ontological resources. These ontological resources also comprise ontology design patterns (Aranguren, 2008; Gangemi, 2005; Blomqvist, 2008; Presutti and Gangemi, 2008), which are generic templates or abstract descriptions proposed to enforce best practices in ontology implementation. One particularity of NeON matching well with our specific approach is that it also takes into account reusing and reengineering of non-ontological resources, which is not the case of other methodologies such as METHONTOLOGY (Fernández-López et al., 1997) and On-To-Knowledge (Sure et al., 2004). These non-ontological resources may consist of structured data such as terminologies (Jimeno-Yepes et al., 2009) or databases, unstructured data (e.g., articles), or semi-structured data (e.g., XML, JSON) (Qawasmeh et al., 2018). In addition to these nine scenarios, NeON also integrates support activities such as knowledge acquisition, documentation, and evaluation that should be carried out during the whole ontology development cycle.

MATERIAL AND METHODS

Summary of the Method
Application of the NeON methodology
We applied the following scenarios of the NeON methodology for implementing OntoADR (**Figure 5**).

- Scenario 1. From Specification to Implementation: this includes four steps (specification, conceptualization, formalization, and implementation). We previously presented these steps in previous work [(Bousquet et al., 2014)] and limit here the

FIGURE 5 | Scenarios of the NeON methodology for implementing OntoADR.
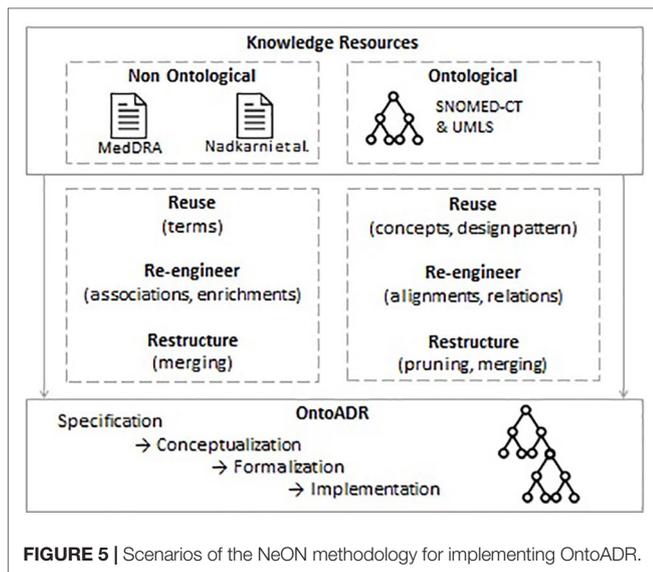
scope of this presentation to scenarios that emphasize the reuse of ontological and non-ontological resources.

- Scenario 2. Reusing and Re-engineering Non-Ontological Resources: We transformed MedDRA into a subsumption tree (see [(Bousquet et al., 2014)], and retrieved a non-ontological resource provided by Nadkarni & Darer (2010). We analyzed this non-ontological resource in order to establish the correspondence between its content (mappings between MedDRA terms and SNOMED CT) and formal definitions that benefit from explicit relations derived from the SNOMED CT concept model. Then, we generated the formal definitions based on this non-ontological resource for inclusion in OntoADR.
- Scenario 6. Reusing, Merging, and Re-engineering Ontological Resources: We reused available ontological resources (UMLS, SNOMED CT) and merged SNOMED CT with MedDRA using mappings available in UMLS. We also reused the SNOMED CT concept model and had to reengineer it to keep only classes and relations that are relevant to formally define MedDRA terms. While the concept model may be considered as a building block that enforces best practices for ontological design in SNOMED CT, it does not strictly correspond to the description of content ontology design patterns, which explains why we did not implement Scenario 7: Reusing ontology design patterns (ODPs).
- Scenario 8. Restructuring Ontological Resources: This consists in pruning parts of the SNOMED CT that are not relevant for the description that was previously described by Souvignet et al., (2016b) and enriching the ontology by adding supplementary concepts and axioms.

In addition to these different scenarios, we also implemented "ontology support activities" for "knowledge acquisition" that comprises activities for (1) capturing knowledge from the MedDRA labels and work by a domain expert for adding formal definition using the Ci4SeR tool [(Souvignet et al., 2014)] and (2) "ontology validation" that consists in checking that the meaning of the ontology definitions are compliant with the definitions we intended the MedDRA terms to convey.

Flow chart of the method

Figure 6 depicts a flow chart representing an overall representation of the several steps and tasks proposed in the article to get an overview of the algorithm at a glance. While this diagram could make readers believe that all these steps were conducted in parallel, it is proposed only as a convenient way to apprehend the method as a whole. The previous paragraph where we applied the NeON methodology shows a different perspective where different scenarios were applied at different time. In the flow chart, each MedDRA term is considered one after the other and can go through several parallel paths according to different conditions.

- If a MedDRA term is associated to a SNOMED CT concept in the UMLS metathesaurus, one mapping is manually selected: the use of the mapping information is described in the section *Using MedDRA-to-SNOMED CT Mappings From UMLS Metathesaurus*.
- If a MedDRA term is present in another mapping source than the UMLS (e.g., the Nadkarni and Darer's proposal), then this mapping is used for the definition in a way that is described in the following section: *Using Another MedDRA-to-SNOMED CT Mapping Resource*.
- If a MedDRA term is composed of several distinct terms by a conjunction (e.g., "acute and chronic thyroiditis" that consists of two medical conditions "acute thyroiditis" and "chronic thyroiditis"), then the MedDRA term is decomposed according to the algorithm described in the section *Using a Syntactic Decomposition Algorithm on Complex MedDRA Terms*. Then, MedDRA subterms are considered as additional MedDRA terms that can be used as a new input for the whole algorithm.
- If a MedDRA term contains a substring with associated meaning (e.g., medical words ending in -algia that indicate pain), this MedDRA term may benefit from a potential syntactic enrichment: the generation of a partial definition is described in the section *Automatic Lexical Enrichment Methods*.

Some manual definitions can optionally be added (see the section *NeON Methodology*). All partial definitions acquired with the different algorithms are then automatically combined into a merged definition of the MedDRA term. We used MedDRA version 17 that consists of 26 SOC, 334 HLGT, 1,720 HLT, 20.559 PT, and 72,637 LLT, SNOMED CT version March 2015 and UMLS version 2014AB. SNOMED CT concepts were extracted from the Concepts_Core_INT file (Release Format 1) and the hierarchy and semantic properties from the Relationships_Core_INT file. This version of MedDRA was applied to the following paths: "using UMLS metathesaurus mappings," "automatic enrichment methods," and "manual definition of concepts." MedDRA 13 that consists of 26 SOC, 335 HLGT, 1,709 HLT, 18,786 PT, and 68,258 LLT was applied to the following paths: "Using other mapping resources" and "Using a decomposition algorithm and Metamap software to map complex MedDRA terms."

Problems With Mapping Other Layers Than the PT Level

We have tried to map other layers (SOC, HLT, and HLGT). For instance, the cardiac disorders SOC concept has for formal definition HASFINDINGSITE some "HEART STRUCTURE." While
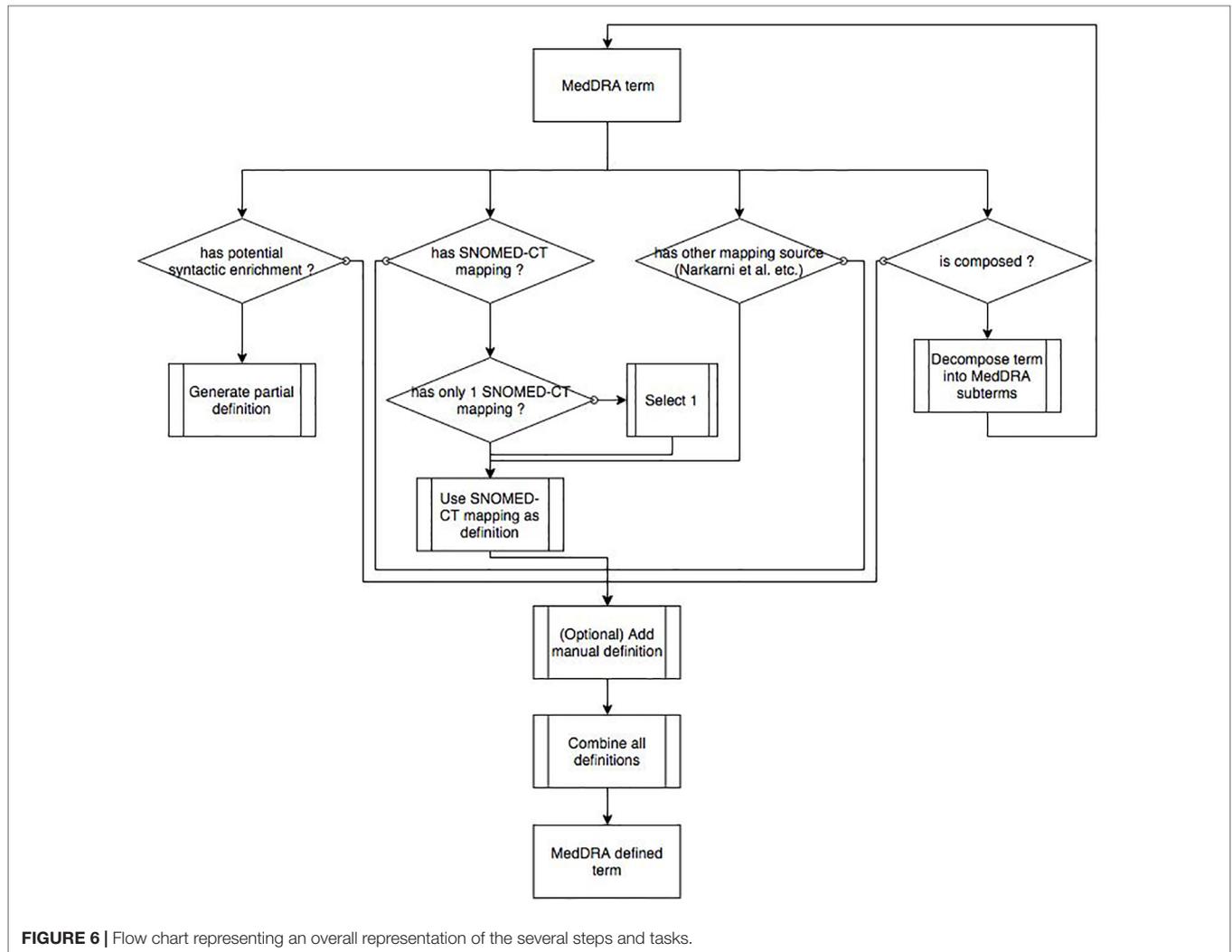
FIGURE 6 | Flow chart representing an overall representation of the several steps and tasks.

some SOC, HLGT, and HLT were accurately defined, essentially thanks to mappings in UMLS, it was decided not to present the results in this article. We used formal definitions associated with the three MedDRA higher levels only in the Ci4SeR tool (Souvignet et al., 2014), where the curator could use them if she considered them as relevant. We decided not to map the LLT because PT is the preferred level for case report analysis and search in pharmacovigilance databases.

We explained in previous work why the MedDRA hierarchy cannot be converted into a subsumption tree because this sometimes causes semantic inconsistencies (Bousquet et al., 2014). The reason is that most high level categories in MedDRA are intended to reflect the domain actors' practices (i.e., following the different medical specialties) and are not necessarily organized according to different semantic criteria as one would expect in a well-formed ontology. A first example concerns groups of symptoms (HLGT or HLT) that are placed under the general categories of disorders that they are the symptom of (SOC or HLGT). Such hierarchical organization would not be authorized in an ontology as the relation being-a-symptom-of does not imply an is-a relation. For instance, the PT "dyspnoea" and "dizziness

or syncope" belong to the HLGT "cardiac disorder signs and symptoms" that is under the SOC cardiac disorders. While dyspnea for instance refers to conditions that may be associated to cardiac disorders, such symptom cannot be considered as a cardiac disorder. A second example is the MedDRA PT sudden death that belongs to the following hierarchies: 1) ventricular arrhythmias and cardiac arrest (HLT)/cardiac arrhythmias (HLGT), and 2) death and sudden death (HLT)/fatal outcomes (HLGT). While "cardiac arrhythmias" is defined in OntoADR with the HASFINDINGSITE SOME "HEART STRUCTURE" property, the sudden death PT should not inherit from such property because sudden death could be the consequence of death that is not of cardiac origin.

Using Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mappings From UMLS Metathesaurus

The UMLS may be used as a source of knowledge for adding formal definitions to medical terminologies (Schulz and Hahn, 2001). Based on our initial experience (Alecu et al., 2008), we assume that SNOMED CT is currently the best candidate for

providing formal definitions to MedDRA. SNOMED CT terms are defined using description logic (DL) formalism, and a fair number of alignments between MedDRA and SNOMED CT are present in the UMLS metathesaurus. Therefore, most of the formal definitions attributed to MedDRA terms in OntoADR are based on semantic information extracted from the SNOMED CT clinical terminology. When a MedDRA term is mapped to a SNOMED CT concept, we reused the semantic information within SNOMED CT in order to build the formal definition of the MedDRA term. Identifying reliable MedDRA-to-SNOMED CT mappings is thus an essential step in our methodology to define MedDRA term semantics.

The UMLS (Lindberg et al., 1993) consists of a semantic network and a metathesaurus developed by the US National Library of Medicine to link terms from more than a hundred controlled vocabularies, including SNOMED CT and MedDRA. Terms from the different vocabularies are linked together by association to a unique UMLS concept defined by a concept unique identifier, e.g., "C0019163" that is mapped to both MedDRA term "hepatitis B" and SNOMED CT concept "type B viral hepatitis."

In OntoADR, MedDRA term "hepatitis B" has the formal definition: hasFindingSite some "Liver Structure," hasAssociatedMorphology some "Inflammation Morphology," and hasCausativeAgent some "Hepatitis B Virus" where hasFindingSite, hasAssociatedMorphology, and hasCausativeAgent are OntoADR semantic relations inspired from SNOMED CT, and "liver structure," "inflammation morphology," and "hepatitis B virus" are SNOMED CT concepts we imported in OntoADR (**Figure 1**).

In order to map MedDRA and SNOMED CT terms, we developed an algorithm following these steps: i) search for the MedDRA PT in the UMLS using the MedDRA identifier; ii) for MedDRA PT without SNOMED CT mappings in the UMLS, if the PT has one or more related LLT considered as synonymous, then the LLT identifier is used for a new UMLS search. iii) If this second search is unsuccessful, the algorithm performs a last UMLS search using PT and LLT labels, seeking to pair these labels with SNOMED CT concepts by string matching.

All mapping propositions selected from the UMLS metathesaurus were validated, modified, or completed by knowledge engineers and pharmacovigilance experts of our team. i) All one-to-one mappings we decided to use were first validated by checking manually the correspondence between the meanings of terms.

Several SNOMED CT concepts may be proposed as synonyms (Fung et al., 2005) of the same MedDRA concept, although they have different meanings. Each MedDRA concept in OntoADR can have only one equivalent SNOMED CT concept. When several SNOMED CT concepts are proposed in UMLS as synonym of a MedDRA term, only one was selected by an expert for building the formal definition. Such selection should be based on synonymy between a SNOMED CT concept and a MedDRA term. According to Fung, synonymy between term X and Y may be defined according to linguistic criteria (Fung et al., 2005) such as enforcing that it is possible to replace X by Y in any sentence without modifying the meaning. Examples of such synonyms are "celiac disease" and "gluten enteropathy," or "kidney stone"

and "renal calculus." Selection of a SNOMED CT concept was performed first by comparing its label with the MedDRA term label. In case both were identical, which occurred most of the time, it was obvious to select this mapping, but in other cases, we took into account the medical relevance of the mapping and had to rely on expert evaluation. For example, three SNOMED CT concepts are mapped to the MedDRA term "Spondylitis": "inflammatory spondylopathy," "undifferentiated spondylitis," and "spondylitis," and the later appears as a perfect match according to label comparison.

Such a validation process was necessary because UMLS mapping propositions are not always semantically valid. MedDRA terms and SNOMED CT concepts mapped together in UMLS can refer to different medical entities, even if they are homonyms. For instance, the MedDRA term "vascular disorders" and its SNOMED CT homonym "vascular disorder" are mapped together in UMLS; however, the former refers to disorders of blood and lymphatic systems and the later only to disorders of blood vessels (lymphatic system disorders are caught by the concept "disorder of lymphatic system" in SNOMED CT). ii) In case of one-to-n mappings, a manual expert choice was made to select the SNOMED CT concept whose definition best fitted the meaning of the correspondent MedDRA concept. iii) When no SNOMED CT concept among the ones suggested by UMLS was satisfactory, the definition of the correspondent MedDRA concept was made manually. iv) Mapping a MedDRA term with a SNOMED CT concept does not ensure that the former gets a complete (or even a satisfying) formal definition of its semantics: the formal definition can be incomplete or even be literally absent: it is common to find SNOMED CT concepts, for instance psychiatric concepts, that have no definitional properties. When necessary, the semantic properties from SNOMED CT attributed to MedDRA concepts through the mapping procedure were thus completed manually by additional assertions.

Using Another Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mapping Resources

To complete the mappings selected from UMLS, we also made use of Nadkarni and Darer's propositions of mappings (Nadkarni and Darer, 2010). Using one year of data (recorded between July 1, 2008 and April 30, 2009) from the US Food and Drug Administration Adverse Events Reporting System (AERS) pharmacovigilance database, the authors identified 3,705 MedDRA PT that collectively accounted for 95% of case reports. The 3,705 selected MedDRA terms correspond to high-frequency terms in the US Food and Drug Administration database and potentially have a great added value. After eliminating terms already mapped to SNOMED CT concepts in UMLS, they attempted to map manually the remaining terms (786 in total) with software assistance. Most of those terms (733) could be mapped by Nadkarni and Darer with SNOMED CT concepts *via* one-to-one or one-to-n mappings.

Several problems have been encountered when trying to reuse Nadkarni and Darer's propositions of mappings

(Nadkarni and Darer, 2010). i) First, in the case of one-to-n mappings, the authors broke down a MedDRA term in such a way to associate it to several SNOMED CT concepts but did not specify which semantic relation was relevant. For example, they mapped the MedDRA concept "tongue discoloration" with SNOMED CT concepts "abnormal color" and "entire tongue" but did not specify which semantic relation interconnected the first to the others. Obviously, it cannot be here an equivalence or synonymy (Fung et al., 2005) ("same as") relation as in the case of one-to-one mapping. When SNOMED CT concepts belong to branches such as *body structure* or *morphologic abnormality*, the relationship to use is easy to deduce, and its creation can be automated: it will be in the first case the HASFINDINGSITE relationship and, in the second, the HASASSOCIATEDMORPHOLOGY relationship. However, when it comes to SNOMED CT concepts from branches *finding*, *qualifier value*, or *disorder*, the relationship to use is not obvious, and only a human expert can decide. A major part of our recovery work was to specify these relationships by making use of the set of relationships available in OntoADR.

ii) We have also occasionally been forced to revise the proposed Nadkarni and Darer's mappings, partly for reasons of pure semantic accuracy and partly because of the purpose of OntoADR, which we illustrate here using three examples:

- For the MedDRA term "feeding disorder neonatal," the authors propose a mapping with SNOMED CT concepts "feeding disorder of infancy OR early childhood" and "neonatal." The second mapping is correct, and we have included "neonatal" in OntoADR to define the MedDRA concept "feeding disorder neonatal" with the HASOCCURRENCE relationship (HASOCCURRENCE SOME NEONATAL). We did not, however, reuse the first mapping. The concept "feeding disorder of infancy OR early childhood" is broader than the MedDRA concept "feeding disorder neonatal," which concerns only the newborn. A mapping with SNOMED CT concept "feeding problems in newborn" would have been more accurate. To complete the formal definition of "feeding disorder neonatal," we added the properties INTERPRETS SOME "FEEDING PATTERN" and HASINTERPRETATION SOME "ABNORMAL."
- The MedDRA concept "azotaemia" is mapped by Nadkarni and Darer to SNOMED CT concepts "blood urea nitrogen measurement" and "increased." We used the first by giving the MedDRA concept "azotaemia" the property INTERPRETS SOME BLOOD UREA NITROGEN MEASUREMENT in OntoADR. However, the use of the second to describe the relationship HASINTERPRETATION appeared problematic. Indeed, azotemia is characterized not so much by an increased concentration of nitrogen compounds in the blood but as a concentration above a certain reference threshold. We thus opted for the creation of the property HASINTERPRETATION SOME ABOVE REFERENCE RANGE, more accurate in this context.
- For the MedDRA term "anorectal discomfort," the authors propose a mapping with SNOMED CT concepts "discomfort" and "anus and rectum (combined site)." However, the problem here is that the SNOMED CT concept "anus and rectum

(combined site)" is set in an isolated portion of the SNOMED CT branch "body structure" (branch called "group of anatomical entities"). Nothing connects it to the concepts of the digestive system structures (e.g., no relationship PART-OF). Due to the SEP decomposition (structure, entire, part) of the anatomical branch of SNOMED CT, the concept "anal structure" has no relation to the concept "anus and rectum." It would have been impossible to use this localization by semantic reasoning, for example, to identify concepts located on part of the anorectal system (principle of subsumption reasoning: concepts that have a relationship of location on parts of the anorectal structure are considered by inference as siblings of the concept of diseases that are located on the whole anorectal structure). We therefore preferred to use the SNOMED CT concept "anorectal structure" to define the relationship HASFINDINGSITE of the MedDRA concept "anorectal discomfort" in OntoADR. This SNOMED CT concept allows the semantic reasoning operation described previously. Moreover, we can assume that the MedDRA term "anorectal discomfort" is sometimes used to encode ADRs in a non-specified way that may be anal or rectal, and not both, as is implied by the use of the SNOMED CT concept "anus and rectum (combined site)." It is therefore important to locate by subsumption concepts that are located within a substructure of the whole anorectal structure.

Using a Syntactic Decomposition Algorithm on Complex Medical Dictionary for Regulatory Activities Terms

Among MedDRA terms that are not mapped with SNOMED CT terms in UMLS, there are many complex terms, i.e., corresponding to composed expressions. MedDRA complex terms are of several kinds: a) expressions composed with an AND logical operator or commas (e.g., "acute and chronic thyroiditis" or "pregnancy, labour, delivery, and postpartum conditions"); b) expressions composed with "NEC" (not elsewhere classified), "unspecified," or with a text between brackets, usually to specify exclusion clauses [e.g., "autoimmune disorders NEC," "laryngeal neoplasms malignancy unspecified," or "ocular neoplasms malignant (excl. melanomas)"]; c) they can also combine these different kinds of complexity [e.g., "gastrointestinal and abdominal pains (excl. oral and throat)" or "ocular structural change, deposit, and degeneration NEC"]. These terms are usually terms of level HLT, HLGT, and SOC in the MedDRA hierarchy. However, their definitions have a great added value because some terms they subsume may inherit their properties. Indeed, defining one high level term with a morphology property may amount to defining all child terms with this property within the limits of what we have indicated in the section *Problems With Mapping Other Layers Than the PT Level*.

The complex MedDRA terms present two kinds of difficulties: i) the difficulty of mapping with SNOMED CT that tends to favor simple concepts probably because most complex concepts correspond to pure classifying artifacts, e.g., "not elsewhere classified," without real counterpart in the phenomena that are part of medicine; ii) difficulties for formalization of meaning: representing

in OWL (Web Ontology Language) the meaning of a compound concept containing exclusions with logical operators is constrained by the expressiveness of the DL language used. In OntoADR, it is not possible to describe the exact same MedDRA semantics due to computability constraints. To date, we have developed a technical solution for the first point, but no satisfactory solution of conceptualization (especially in terms of human cost modeling) has yet been developed to meet the second point. It should be noted that this issue is regarding mainly terms of high levels and does not affect the progress of definitions for PT terms in OntoADR.

In order to map complex MedDRA terms, we developed an algorithm for syntactic decomposition. It consists of three routines: 1) a routine for "cleaning" terms; 2) a routine for identification of complex expressions; and 3) a routine for decomposition of an expression from a set of formal rules. Routine 1 begins by suppressing from the MedDRA labels unnecessary characters or characters that cannot be supported by the decomposition routine [stop words, content between brackets, terms as "unspecified," "NOS" (not otherwise specified), etc.]. Routine 2 identifies decomposable expressions: it searches for keywords that indicate a probable composition of the expression ("AND," "OR," "WITH," ";", etc.). Finally, routine 3 decomposes the complex expression in a set of simpler expressions (cf. **Table 2**), by applying different rules, for example:

$$(A \text{ AND } B).q \rightarrow A.q + B.q$$
$$q.(A \text{ AND } B) \rightarrow q.A + q.B$$

We then used the MetaMap software (Aronson, 2001) to map all new decomposed concepts to existing SNOMED CT concepts.

Automatic Lexical Enrichment Methods

We have used a rule-based algorithm for automatic suggestion of properties from the MedDRA label to enrich the formal definition of concepts. Two key procedures have been implemented in the algorithm:

1. When the algorithm detects a given string S_x in a MedDRA label, it automatically adds a corresponding property P_x in the OWL concept definition. For example: if the string "pain" or "algia" is found in a MedDRA concept's label, the semantic property HASDEFINITIONALMANIFESTATION SOME PAIN is automatically added to the concept's definition. Similarly, if the string "perforation" is found, the formal definition HASASSOCIATEDMORPHOLOGY SOME PERFORATION is suggested. All created properties are then validated by an expert. Illegitimate properties are rejected. For example, the algorithm proposed to add the formal property HASASSOCIATEDMORPHOLOGY SOME HERNIA to the MedDRA concept "hernia repair," as the string "hernia" was found in the label. Semantically, this assignment is obviously illegitimate: the "hernia repair" is not a type of hernia and cannot be defined by this morphological property. The property has therefore been rejected. The expert, however, took advantage of this suggestion to correct it in: OCCURSAFTER

TABLE 2 | Examples of parsing of complex MedDRA terms using different rules.

INPUT : MedDRA complex term	Selected Rule	OUTPUT : result of the decomposition
"Ear and labyrinth disorders"	(A AND B).q	"ear disorders" "labyrinth disorders"
"Manic and bipolar mood disorders and disturbances"	((A AND B).q).(C AND D)	"manic mood disorders" "bipolar mood disorders" "manic mood disturbances" "bipolar mood disturbances"
"Blood and lymphatic system disorders"	(A AND B).q	"blood system disorders" "lymphatic system disorders"

SOME HERNIA. This validation step is also necessary due to the occasionally polysemic expressions used for automatic generation of properties. For example, the automatic generation of the HASCLINICALCOURSE SOME CYCLIC property, when the algorithm detects the "cyclic" string in a MedDRA label is valid for concepts such as "cyclic neutropenia" or "cyclic vomiting syndrome," where the term "cyclic" indicates the clinical course of the disease. However, it is not valid for the concept "cyclic AMP," which refers to a clinical test (a measure of the presence or amount of cyclic adenosine monophosphate, e.g., in urine). We could have improved the automatic processing in order to detect these problematic cases, but the formalization of these exceptions would have taken longer time than using a manual approval process.

A restriction is applied to prevent the property P_x to be duplicated when it already exists in a MedDRA term definition, e.g., the detection of the "perforation" string in the label of a MedDRA concept C_{Med} only results in the creation of the property HASASSOCIATEDMORPHOLOGY SOME PERFORATION if C_{Med} does not already own a property HASASSOCIATEDMORPHOLOGY SOME <MORPHOLOGY>. If it is the case, we assume that the relation R_x (in this case HASASSOCIATEDMORPHOLOGY) has already been filled in correctly.

2. A second procedure is implemented by the algorithm to automatically generate properties. Based on the same principles, but working with more complex patterns of recognition, it was designed to complete definitions of MedDRA concepts referring to investigations and their results (SOC « Investigations »).

Two relationships are available in SNOMED CT to define the examination results (whether clinical observations or investigations): INTERPRETS, which refers to "the entity being evaluated or interpreted, when an evaluation, interpretation, or "judgment" is intrinsic to the meaning of a concept"; and HASINTERPRETATION, which, grouped with the attribute INTERPRETS, "designates the judgment aspect being evaluated or interpreted for a concept (e.g., presence, absence, degree, normality, abnormality, etc.)" (Rector and Brandt, 2008). It is important that these two relationships are filled in OntoADR in order to apply semantic reasoning not only to ADR concepts as such, but also to concepts referring to abnormal results of investigations that are the consequence of an ADR (for instance,

"neutrophil count decreased" for the neutropenia condition), as such results are frequently used to describe ADRs in pharmacovigilance databases. However, it turned out that very few MedDRA concepts located in the investigations branch could be identified through the procedures described in the previous sections, in particular the mapping from UMLS. A large majority of MedDRA concepts in SOC investigations thus remained undefined in OntoADR.

To remedy this situation, we have integrated into the algorithm a module supporting the properties INTERPRETS and HASINTERPRETATION for MedDRA concepts from SOC "investigations." Results of investigations are usually expressed in MedDRA using the following adjectives: abnormal, normal, absent, present, increased, decreased, positive, and negative. All these qualifiers are also used in SNOMED CT to fill the property HASINTERPRETATION. The procedure followed by the algorithm was therefore as follows:

When the string S_x corresponding to one of these adjectives is detected in the label < LAB1> of a MedDRA concept C_{Med1} from SOC "Investigations":

1. Create in the definition of C_{Med1} the property HASINTERPRETATION SOME S_x.
2. Find if it exists in the investigations branch a concept C_{Med2}, whose label < LAB2> corresponds to (< LAB1> minus S_x). If C_{Med2} exists, create in the definition of C_{Med1} the property INTERPRETS SOME C_{MED2}. This second phase of the procedure is used to connect *via* the property INTERPRETS the results to the related investigations.

In the example of the concept C_{Med1} « Alpha hydroxybutyrate dehydrogenase decreased », this procedure gives the following results:

1. Creation of the property HASINTERPRETATION SOME DECREASED in the definition of C_{Med1}
2. There is a concept C_{Med2}: "alpha hydroxybutyrate dehydrogenase". The property INTERPRETS SOME 'ALPHA HYDROXYBUTYRATE DEHYDROGENASE' is thus created in the definition of C_{Med1}.

Once again, all of the created properties were reviewed and validated by an expert.

Manual Definition

Besides these semiautomatic methods for defining MedDRA concepts in OntoADR, we also performed the manual definition of about 1,935 concepts (Souvignet et al., 2016b). We had insufficient human resources to carry out the manual definition of all MedDRA terms that previous methods had failed to define. So, we decided to focus on high value-added terms for pharmacovigilance. In the EU-ADR project, Trifirò et al. (2009) developed a ranked list of 23 first importance adverse drug events (e.g., cardiac valve fibrosis) based on a review of scientific literature, medical textbooks, and websites of regulatory agencies. To identify which MedDRA terms are related to those 23 topics, pharmacovigilance experts familiar with MedDRA have chosen for each topic an SMQ and/or MedDRA hierarchy-based grouping (HLT or HLGT) or a custom set of

preferred terms (PT) fitting the definition of the targeted topics (see Declerck et al., 2012 for details). When no existing MedDRA groupings could be identified to fit the safety topic, *ad hoc* manual groupings of MedDRA PT were proposed by the experts. This work benefited from using a dedicated tool we implemented, Ci4SeR (curation interface for semantic resources) (Souvignet et al., 2014).

RESULTS

Using Other Medical Dictionary for Regulatory Activities-to-SNOMED Clinical Terms Mapping Resources

"Once the Nadkarni and Darer's mapping propositions were validated, modified or completed, we applied the same procedure as described in the section *Using MedDRA-to-SNOMED CT Mappings From UMLS Metathesaurus* to pick up information from SNOMED CT and define the MedDRA concepts of OntoADR. Using the set of SNOMED CT relations available in OntoADR, we also realized manually the definition of those MedDRA terms (53 in total) for which no mapping could be found by Nadkarni and Darer. The use, after verification and eventually correction and complementation, of mappings proposed by Nadkarni and Darer, allowed us to complete the definition of 786 supplementary MedDRA PTs in OntoADR.

Using a Syntactic Decomposition Algorithm on Complex Medical Dictionary for Regulatory Activities Terms

Among the 2,070 HLT, HLGT, and SOC in MedDRA 13.0, a total of 1,011 terms was decomposed by the algorithm generating an average of 2.7 terms by decomposition. The consistency of automatic decomposition was checked by an expert. The errors were corrected through a progressive adjustment of the decomposition algorithm. Only the decomposition of 30 complex terms that were not supported by the algorithm was done manually. Once the decomposition was performed, we used the UMLS MetaMap 2010 AB mapping software, which returns from a given string (in our case, a part of the decomposition), the UMLS concept unique identifier of the nearest syntactically SNOMED CT concepts (fuzzy match). With this method, a total of 638 MedDRA concepts (9 SOCs, 131 HLGTs, and 498 HLTs) could be mapped to the SNOMED CT concepts (mappings one-to-one or one-to-n).

This additional mapping method has the advantage of enabling the definition of high level terms in MedDRA. These definitions may then be inherited by subsumed low level terms. However, the definitions have also the disadvantage of being broad and thus potentially insufficiently precise for specific preferred terms.

Automatic Lexical Enrichment Methods

This procedure was applied to 11 of the 25 SNOMED CT properties used in OntoADR, using 82 different matching strings. In total, this procedure has led to the creation of 8,194 properties, among which 7,691 were validated (i.e.,

93.9%). A sample of the strings detected by the algorithm and properties created is shown in **Table 3**.

Manual Definition of Concepts

Figure 7 depicts as an example the formal definition associated to the term "Shwachman-Diamond syndrome," as it was described in OntoADR after application of the different algorithms that precede manual refinement.

The curation, which took approximately 750 h, allowed refining the definition of 1,935 MedDRA terms to validate and fully define these terms (Souvignet et al., 2016b). Among the 3,482 properties available in OntoADR for these terms, the curator validated 2,636 properties (76%), proposed 350 (10%) more precise terms (i.e., narrower terms in the SNOMED CT hierarchy), and removed 496 properties (14%). The curator also proposed 13,675 additional properties, but these should not be considered as errors related to missing properties but rather

TABLE 3 | Sample of the properties created automatically from the MedDRA label to enrich the formal definitions of MedDRA concepts in OntoADR.

Relation	Matching strings	Value of the property	Nb properties created	% properties validated
hasClinicalCourse	acute	Sudden onset and/or short duration	84	95.1%
	cyclic	Cyclic	10	20%
	recurrent	Recurrent	121	100%
hasCausativeAgent	bacteria	Bacteria	83	100%
	viral	Virus	157	88.5%
hasAssociatedMorphology	abscess	Abscess morphology	121	100%
	hernia	Hernia	70	82.9%
	haemorrhage, haemorrhagic, bleeding	Hemorrhage	272	86.4%
hasPathologicalProcess	infection, infections, infectious,	Infectious process	726	89.4%
	infective parasitic	Parasitic process	12	91.7%
Interprets	motor, movement, kines	Motor function behaviour	66	59.1%
DueTo	allergic	Hypersensitive reaction	32	93.8%

FIGURE 7 | Formal definition associated to the preferred term "Shwachman-Diamond syndrome" before manual refinement.

as the curator's desire to better document diagnoses with signs and symptoms and investigations that may be associated to a given disease but are not specific, as they may be absent in some occurrences of this disease.

Figure 8 shows how the "Shwachman-Diamond syndrome" PT's formal definition was modified by the curator in the Ci4SeR tool. The lowest part of the screenshot contains the properties that were automatically proposed considering the parent's and siblings' formal definitions. **Table 4** depicts the results using each method.

DISCUSSION

Summary

We have described in this article several methods that allow collectively a better semantic enrichment of MedDRA. **Table 4** shows that using UMLS metathesaurus is the method that was the most efficient considering the number of mappings and helped to add formal definitions for about half MedDRA terms. As other mapping resources than UMLS are rare and concern
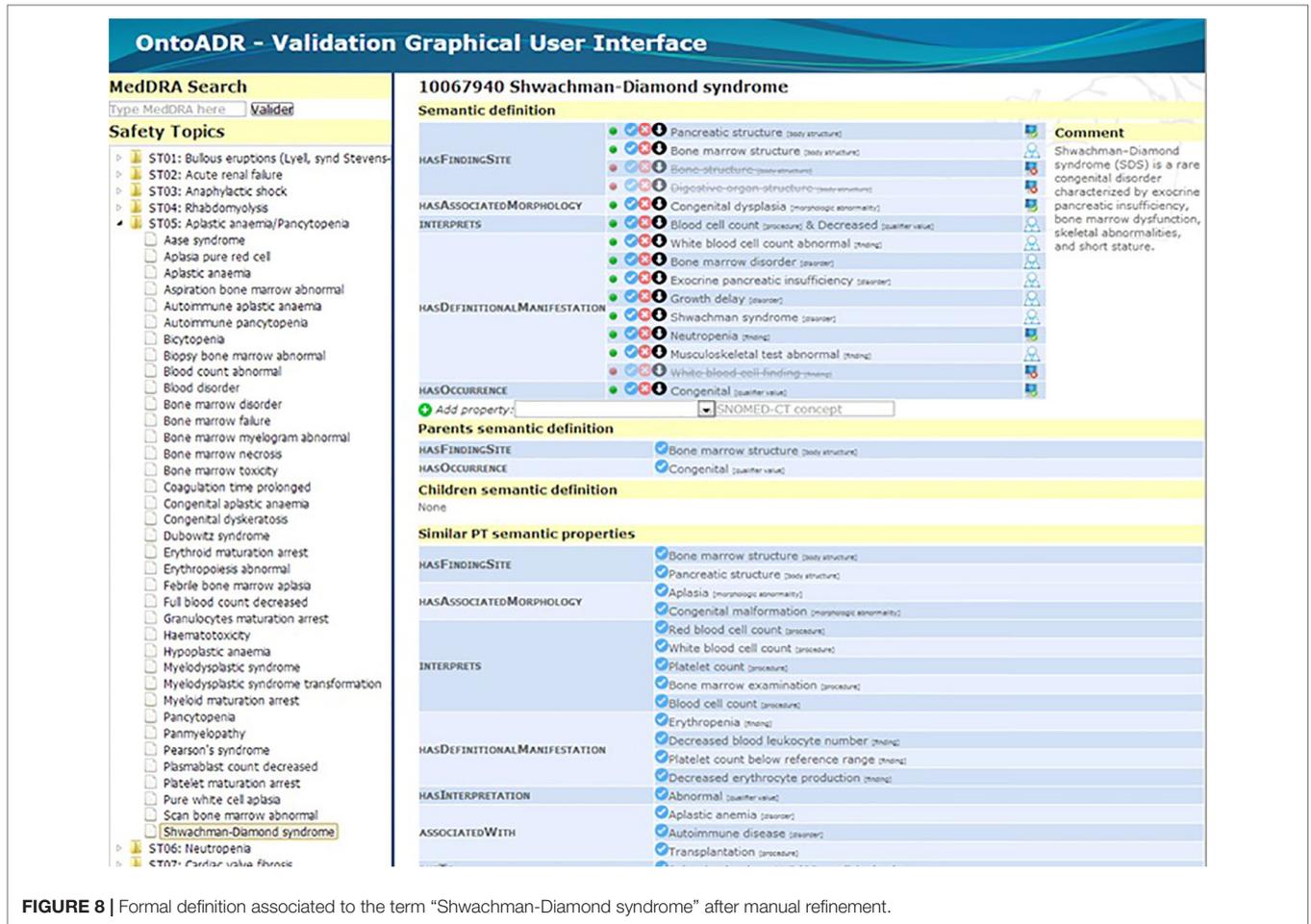

FIGURE 8 | Formal definition associated to the term "Shwachman-Diamond syndrome" after manual refinement.

TABLE 4 | Synthesis of mappings and properties found using all previously described methods.

Source/Method	MedDRA version	Comparator	Number of mappings	Number of properties
Using UMLS Metathesaurus mappings	v17	20,599 PT	11,281 PT (54.8%)	74,598
Using other mapping resources	v13	18,786 PT	455[a] PT (2.4%)	469[a]
Using a decomposition algorithm and Metamap software to map complex MedDRA terms	v13	2,070 HLT, HLGT, and SOC	638[a,b] HLT, HLGT, and SOC (30.8%)	–
Automatic enrichment methods	v17	–	–	7,691[a]
Manual definition of concepts	v17	20,599 PT	1,935 PT (9.1%)	13,675

[a]Represent only the number of mapping/properties that was not found by other methods.
[b]Limited to SOC, HLGT, and HLT.

TABLE 5 | Summary of inconveniences and advantages of the different methods.

Algorithms	What is already available?	Characteristics of the algorithm
Using MedDRA-to-SNOMED CT mappings from UMLS Metathesaurus	MedDRA-to-SNOMED CT mappings in UMLS Metathesaurus	Available mappings are used to retrieve SNOMED CT concepts associated to a MedDRA term. Properties are added to the formal definition according to the SNOMED CT concept position in the hierarchy. It illustrates Scenario 6. "Reusing, Merging and Re-engineering Ontological Resources" of the NeON methodology. The algorithm is automatic, except when several SNOMED CT concepts are available which requires expert selection.
Using other MedDRA-to-SNOMED CT mapping resources	Nadkarni and Darer's propositions of mappings	This is an expert-based process entirely manual of validation and refinement that illustrates Scenario 2. "Reusing and Re-engineering Non-Ontological Resources" of the NeON methodology but benefits from a non-ontological resource that expedites formal definition of MedDRA terms compared with manual definitions.
Using a syntactic decomposition algorithm on complex MedDRA terms	–	This algorithm is automated and developed ad hoc. It illustrates one of the ontology support activities, "knowledge acquisition," and exploits hidden semantics as proposed by Third (2012). It is limited to complex MedDRA terms.
Automatic lexical enrichment methods	–	This algorithm is based on substring search. It necessitates defining beforehand substrings that may be associated to a SNOMED CT concept. Review of the algorithm's proposal is mandatory in order to check that substrings allow associating with relevant SNOMED CT concepts. It also illustrates Ontology support activity "knowledge acquisition" and exploits hidden semantics.
Manual definition	–	This process is manual and expert based, but the Ci4SeR tool suggests definitions on the basis of definitions already available for siblings and parents of a MedDRA PT. Such approach addresses both "knowledge acquisition" and "ontology validation" within the Ontology support activities.

only few MedDRA terms, the Nadkarni and Darer's resource allowed to add properties to 4.2% of MedDRA terms but only to 2.4% of MedDRA terms that were not associated with mappings to SNOMED CT in UMLS.

Our proposal to decompose complex MedDRA terms was applied only to SOC, HLGT, and HLT levels and accounted for 30.8% of these MedDRA terms above the PT level. Manual definitions and refinements of definitions obtained with other methods allowed to process 9.1% of MedDRA terms, which is more than the proportion of terms that were defined using Nadkarni and Darer's mapping resource. However, it was associated with high time-consuming effort by the domain expert that confirms previous work, e.g., Giannangelo and Millar (2012) who observed that "map specialists on average mapped 6.5 SNOMED CT concepts an hour." **Table 5** summarizes the main characteristics of each method and indicates if the proposed method reuses existing knowledge, if it requires manual adaptations or may be performed in an automated way.

Related Work in Medical Informatics

He et al. (2014) have introduced the Ontology of Adverse Events (OAE). OAE was originally targeted for vaccine adverse events (Marcos et al., 2013) and now also includes adverse drug events. In practice, using OAE to select case reports in the Vaccine Adverse Event Reporting System proved difficult: "AE data stored in Vaccine Adverse Event Reporting System are annotated using MedDRA" (Marcos et al., 2013). Authors complained that "many disadvantages of MedDRA, including the lack of term definitions and a well-defined hierarchical and logical structure, prevent its effective usage in VAE (vaccine adverse event) term classification." Therefore, for an efficient analysis, they performed a mapping between MedDRA and OAE (Sarntivijai et al., 2012).

OAE contains about 2,300 AE entities but only 1,900 MedDRA mappings (9% of all MedDRA PT). For example, there is a single term for upper gastrointestinal hemorrhage in OAE (He et al., 2014), whereas one can cite several in MedDRA (see the section *Rationale for Supplementing MedDRA With Formal Definitions* where we identified 27 using OntoADR). Furthermore, OAE formal definitions are limited to anatomical and physiopathological descriptions. He and colleagues proposed extensions to OAE such as the Ontology of Drug Neuropathy Adverse Events (Guo et al., 2016), which suggests that providing supplementary MedDRA mappings is possible using the same methodology. One advantage of OAE is the possibility to use it in open access, which allows wide dissemination to users, while legal issues related to ownership of MedDRA and SNOMED CT should be solved before we can make OntoADR available.

Adverse Events Reporting Ontology aims to allow storing of pharmacovigilance data related to anaphylaxis according to guidelines defined by the Brighton collaboration (Courtot et al., 2014) but may also be extended to other safety topics, e.g., malaria (Courtot et al., 2013). Nevertheless, ADRs are not formally defined in Adverse Events Reporting Ontology.

While we did not find any resource available providing definitions for every ADR in MedDRA, there are more general resources with formal representation of clinical terms. In order not to start from scratch the definitions of ADRs, we needed a trustworthy formal resource, standardized and reliable. We chose SNOMED CT for three main reasons: first, pharmacovigilance concepts generally do not differ from those used in other medical fields. Second, SNOMED CT is the most complete and most detailed terminology of medicine with a formal semantic foundation currently available (Elkin et al., 2006) sharing common fields with MedDRA (medical pathologies in all medical specialties, signs and symptoms, laboratory tests results, some diagnostic and therapeutic procedures). Finally, SNOMED CT has the advantage of covering to a large extent, if not entirely, other standard medical terminologies such as International Classification of Diseases, 10th edition (ICD-10), and especially more than 50% of MedDRA terms (excluding LLT) are associated with a SNOMED CT concept

(Bodenreider, 2009) in UMLS, a degree of coverage that, to our knowledge, no other current medical ontology was able to match.

We found in the literature several examples of mappings from a terminology to SNOMED CT (Vikström et al., 2007; Merabti et al., 2009; Nyström et al., 2010; Dhombres and Bodenreider, 2016; Fung et al., 2017). However, the objective was usually to integrate a terminology in SNOMED CT or to map this terminology to SNOMED CT but not to enrich this terminology by the means of formal definitions. The lexically assign logically refine method is an example of an automated method in which logical observation identifiers names and codes (LOINC) and SNOMED terms are first decomposed, then refined by the means of knowledge-based methods that allowed to map LOINC and SNOMED together (Dolin et al., 1998). In another work, Adamusiak and Adamusiak and Bodenreider (2012) developed an OWL version of both LOINC and SNOMED CT and made use of mappings between SNOMED CT terms to identify redundancy and inconsistencies in LOINC multi-axial hierarchy. Roldán-García et al. (2016) implemented Dione, an OWL representation of ICD-10-CM where formal definitions were obtained thanks to mappings between ICD-10-CM and SNOMED CT available in UMLS and the Bioportal. More recently, Nikiema et al. (2017) benefited from SNOMED CT logical definitions to find mappings between ICD-10 and ICD-O3 concepts in the domain of cancer diagnosis terminologies.

It is usually recommended to build medical terminologies following the model of clinical terminologies that obey to Cimino's desiderata (Cimino, 1998; Bales et al., 2006). Such model brings several advantages such as improving the maintenance of large terminologies (Cimino et al., 1994), and formal definitions were implemented in several terminologies such as the NCI-Thesaurus (Hartel et al., 2005). Our approach is more in line with what is recommended by Ingenerf and Giere (1998), that is to say, to keep terminologies with disjoint classes required for statistics (in a clinical terminology, the same term may be present in several separate categories because of multiple inheritance and be counted more than once) and instead implement a mapping of terms of first-generation system to a formal system. This allows keeping the MedDRA terminology in its current format, counting ADRs according to predefined categories that are standardized and replicable at the international level with MedDRA and building new categories on demand by using knowledge engineering methods. This is what we have done in our implementation of OntoADR (Bousquet et al., 2014) in the form of an OWL-DL file and in the form of a database (Souvignet et al., 2016b).

We have no knowledge of other works in which the formalization of complex terms involving AND/OR relations has been performed in an automated way. We have not proposed formal definitions of LLT because this level is reserved for the coding of case reports, in order to improve the accuracy of coding, but it is not useful for grouping data for analysis (which is performed at the PT level). Although the analysis of pharmacovigilance databases is performed preferentially at the PT level, it could be important to also define the upper levels: SOC, HLGT, and HLT. This formalization would bring several advantages: i) preferred terms may inherit properties from their parents that allows to give them a formal definition in case the

synonymous SNOMED CT concept has no definition, or there is no SNOMED CT concept mapped to this PT in UMLS; ii) This would allow to calculate by the means of terminological reasoning high level MedDRA categories in which PTs should be included and therefore restore multiple inheritance that does not exist in MedDRA. However, it is advisable to remain modest insofar as the relations between a PT and the higher hierarchical levels to which it is attached are not always of a taxonomic nature.

Perspectives

Our perspectives are to add formal definitions to a larger number of MedDRA terms. Our approach may be improved using more advanced natural language processing techniques (Iavindrasana et al., 2006; Deléger et al., 2009; Liu et al., 2011; Dupuch et al., 2014) compared with the basic semantic enrichment we performed considering MedDRA labels. We estimate that the methods proposed here can be reused for other first-generation terminologies provided that these terminologies have a mapping with SNOMED CT with fair coverage and that this mapping is available in accessible sources of knowledge such as the UMLS. The terminology can also be treated using methods of natural language processing as was done for example with LOINC in the lexically assign logically refine method (Dolin et al., 1998). One can also consider cases in which the terminology would be normally defined by mapping to another clinical terminology than SNOMED CT. This may be the case in other areas of application in which SNOMED CT is not the best choice.

As the manual approach was time consuming and necessitates human resources we do not have, we plan to rely on the development of complementary automated approaches. First, formal definitions could be extracted from textual definitions (Petrova et al., 2015) or directly using morphosemantic analysis on the term label, e.g., blepharitis where "itis" stands for "inflammation," and "blephar" stands for "eyelid." Such approach is limited to terms containing "compound forms" that have a medical meaning (Deléger et al., 2009). Second, formal definitions could be based on ontology design patterns, such as implemented in tools like Ontorat (Xiang et al., 2015) or TermGenie (Dietze et al., 2014), which partially automate the process, as they still rely on expert curation. Third, additional mappings between MedDRA and other terminologies could be obtained *via* improved mappings in the UMLS metathesaurus (Bodenreider et al., 1998; Fung et al., 2007; Diallo, 2014). Fourth, semantic definitions may be audited by comparing definitions associated to terms that present lexical similarities (Agrawal and Elhanan, 2014). However, this presents an intrinsic limit: terms to compare should consist of at least three words that constraints this method mainly to MedDRA procedures.

Fifth, we plan to extract knowledge using additional sources than SNOMED CT such as NCI Thesaurus (Sioutos et al., 2007) that could be useful to build definitions for MedDRA terms that describe cancer-related adverse reactions. A recent work by Oliveira and Pesquita, (2018) reports that current ontology matching techniques and systems are mostly devoted to finding links between two equivalent entities from two distinct ontologies. However, different domains may be involved that requires the implementation of matching techniques that

allow linking more than two ontologies through more complex relations. An example is "aortic valve stenosis" (from human phenotype ontology) that is equivalent to the combination of "aortic valve" (from the Foundational Model of Anatomy) and "constricted" (from Phenotype And Trait Ontology).

CONCLUSION

The possibility of selecting terms using formal definitions and terminological reasoning are major advantages of clinical terminologies with formal semantics such as SNOMED CT, which present several advantages compared with classic terminologies. MedDRA, as a standard international terminology for the coding of ADRs in pharmacovigilance databases, could beneficiate from these knowledge engineering techniques, but MedDRA terms have to be defined using formal languages first. As defining manually MedDRA terms takes much time, it is important to reuse as much as possible ontological and non-ontological resources available to expedite the generation of formal definitions. The collection of methods we present can collectively support a semiautomatic semantic enrichment of MedDRA. Perspectives are to implement more efficient techniques to find more logical relations between SNOMED CT and MedDRA in an automated way.

AUTHOR CONTRIBUTIONS

GD adapted Nadkarni and Darer's definitions to OntoADR, performed automatic lexical enrichment methods, and wrote the first draft of the manuscript. M-CJ provided significant advice on the design of the study and contributed to the evaluation. ES first, then JS, performed mappings between MedDRA and SNOMED CT using the UMLS metathesaurus and developed new versions of OntoADR. JS performed the syntactic decomposition algorithm on complex MedDRA terms and wrote the corresponding section of the manuscript. CB conducted the study, contributed to the evaluation, reviewed state-of-the-art related work, and wrote the final article. CB and M-CJ were responsible for submitting the PROTECT project to the IMI requests for proposal and for submitting the PEGASE project to the ANR request for proposal. All authors have made substantial contributions and approved the final manuscript and agreed to be accountable for all aspects of the work.

REFERENCES

Adamusiak, T., and Bodenreider, O. (2012). Quality assurance in LOINC using Description Logic. *AMIA Annu. Symp. Proc.* 2012, 1099–1108.

Agrawal, A., and Elhanan, G. (2014). Contrasting lexical similarity and formal definitions in SNOMED CT: consistency and implications. *J. Biomed. Inform.* 47, 192–198. doi: 10.1016/j.jbi.2013.11.003

Alani, H. (2006). Ontology construction from online ontologies. In: 15th International World Wide Web Conference, 23-26 May 2006, Edinburgh, Scotland. 491–495. doi: 10.1145/1135777.1135849

Alecu, I., Bousquet, C., Mougin, F., and Jaulent, M. C. (2006). Mapping of the WHO-ART terminology on Snomed CT to improve grouping of related adverse drug reactions. *Stud. Health Technol. Inform.* 124, 833–838. doi: 10.3233/978-1-58603-647-8-833

Alecu, I., Bousquet, C., Degoulet, P., and Jaulent, M. C. (2007). PharmARTS: terminology web services for drug safety data coding and retrieval. *Stud. Health Technol. Inform.* 129 (Pt 1), 699–704. doi: 10.3233/978-1-58603-774-1-699

Alecu, I., Bousquet, C., and Jaulent, M. C. (2008). A case report: using SNOMED CT for grouping Adverse Drug Reactions Terms. *BMC Med. Inform. Decis. Mak.* 8Suppl 1, S4. doi: 10.1186/1472-6947-8-S1-S4

Alobaidi, M., Malik, K. M., and Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics* 19 (1), 319. doi: 10.1186/s12859-018-2339-3

Aranguren, M. E., Antezana, E., Kuiper, M., and Stevens, R. (2008). Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. *BMC Bioinformatics.* 9 Suppl 5:S1. doi: 10.1186/1471-2105-9-S5-S1

Aronson, A. R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in *Proceedings of the AMIA Symposium* (American Medical Informatics Association), 17.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., and Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018, bay101. doi: 10.1093/database/bay101

Asfari, H., Souvignet, J., Lillo-Le Louët, A., Trombert, B., Jaulent, M. C., and Bousquet, C. (2016). Automated grouping of terms associated to cardiac valve fibrosis in MedDRA. *Therapie.* 71 (6), 541–552. doi: 10.1016/j.therap.2016.06.003

Astrakhantsev, N. A., and Turdakov, D. Y. (2013). Automatic construction and enrichment of informal ontologies: A survey. *Program Com. Software* 39 (1), 34–42. doi: 10.1134/S0361768813010039

Balakrishna, M., Moldovan, D. I., Tatu, M., and Olteanu, M. (2010). Semi-automatic domain ontology creation from text resources In: Proc. 7th International Conference on Language Resources and Evaluation (LREC 10). Valletta, Malta, 19–21.

Bales, M. E., Kukafka, R., Burkhardt, A., and Friedman, C. (2006). Qualitative assessment of the International Classification of Functioning, Disability, and Health with respect to the desiderata for controlled medical vocabularies. *Int. J. Med. Inform.* 75 (5), 384–395. doi: 10.1016/j.ijmedinf.2005.07.026

Bedini, I., and Nguyen, B. (2007). *Automatic ontology generation: State of the art. PRiSM Laboratory Technical Report*. University of Versailles.

Blfgeh, A., Warrender, J., Hilkens, C. M. U., and Lord, P. (2017). A document-centric approach for developing the tolAPC ontology. *J. Biomed. Semantics* 8 (1), 54. doi: 10.1186/s13326-017-0159-4

Blomqvist, E. (2008). Pattern ranking for semi-automatic ontology construction. In Proceedings of the 2008 ACM symposium on Applied computing pp. 2248–2255.

Bobed, C., Mena, E., and Trillo, R., (2012). "FirstOnt: Automatic Construction of Ontologies out of Multiple Ontological Resources," in *Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (Spain: KES'12, IOS Press), 1909–1919.

Bodenreider, O., Nelson, S. J., Hole, W. T., and Chang, H. F., (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings/AMIA Annual Symposium*. 815–819.

Bodenreider, O. (2009). Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. *AMIA. Annu. Symp. Proc.* 14, 45–49.

Bousquet, C., Lagier, G., Lillo-Le Louët, A., Le Beller, C., Venot, A., and Jaulent, M. C. (2005). Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions. *Drug Saf.* 28 (1), 19–34. doi: 10.2165/00002018-200528010-00002

Bousquet, C., Sadou, E., Souvignet, J., Jaulent, M. C., and Declerck, G. (2014). Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms. *J. Biomed. Inform.* 49,282–291 pii: S1532–0464(14)00079-3. doi: 10.1016/j.jbi.2014.03.012

Brown, E. G., Wood, L., and Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* 20 (2), 109–117. doi: 10.2165/00002018-199920020-00002

Brown, E. G. (2003). Methods and pitfalls in searching drug safety databases utilising the Medical Dictionary for Regulatory Activities (MedDRA). *Drug Saf.* 26 (3), 145–158. doi: 10.2165/00002018-200326030-00002

Buitelaar, P., Cimiano, P., and Magnini, B., (2005). Ontology learning from text: an overview. Ontology learning from text: methods, evaluation and applications 123, 3–12.

Cimino, J. J., Clayton, P. D., Hripcsak, G., and Johnson, S. B. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J. Am. Med. Inform. Assoc.* 1 (1), 35–50. doi: 10.1136/jamia.1994.95236135

Cimiano, P., Völker, J., and Studer, R. (2006). Ontologies on demand?-a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. *Information, Wissenschaft und Praxis* 57 (6-7), 315–320.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf. Med.* 37 (4-5), 394–403. doi: 10.1055/s-0038-1634558

Costa, R., Lima, C., Sarraipa, J., and Jardim-Gonçalves, R. (2016). Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach. *J. Intell. Manuf.* 27 (1), 263–282. doi: 10.1007/s10845-013-0856-5

Courtot, M., Zheng, J., Stoeckert, C. J., Jr., Brinkman, R. R., and Ruttenberg, A. (2013). Diagnostic criteria and clinical guidelines standardization to automate case classification. *In ICBO*, 101–104.

Courtot, M., Brinkman, R. R., and Ruttenberg, A. (2014). The logic of surveillance guidelines: an analysis of vaccine adverse event reports from an ontological perspective. *PLoS One* 9 (3), e92632. doi: 10.1371/journal.pone.0092632

Cullen, J., and Bryman, A. (1988). The knowledge acquisition bottleneck: time for reassessment? *Expert Syst.* 5 (3), 216–225. doi: 10.1111/j.1468-0394.1988.tb00065.x

Dahab, M. Y., Hassan, H. A., and Rafea, A. (2008). TextOntoEx: automatic ontology construction from natural English text. *Expert Syst. Appl.* 34 (2), 1474–1480. doi: 10.1016/j.eswa.2007.01.043

Dasgupta, S., Padia, A., Maheshwari, G., Trivedi, P., and Lehmann, J., (2018). Formal Ontology Learning from English IS-A Sentences. *arXiv preprint arXiv*:1802.03701.

Declerck, G., Bousquet, C., and Jaulent, M. C. (2012). Automatic generation of MedDRA terms groupings using an ontology. *Stud. Health Technol. Inform.* 180, 73–77. doi: 10.3233/978-1-61499-101-4-73

Deléger, L., Namer, F., and Zweigenbaum, P. (2009). Morphosemantic parsing of medical compound words: transferring a French analyzer to English. *Int. J. Med. Inform.* 78 Suppl 1, S48–S55. doi: 10.1016/j.ijmedinf.2008.07.016

Dhombres, F., and Bodenreider, O. (2016). Interoperability between phenotypes in research and healthcare terminologies–Investigating partial mappings between HPO and SNOMED CT. *J. Biomed. Semantics* 7, 3. doi: 10.1186/s13326-016-0047-3

Diallo, G. (2014). An effective method of large scale ontology matching. *J. Biomed. Semantics* 5 (1), 44. doi: 10.1186/2041-1480-5-44

Dietze, H., Berardini, T. Z., Foulger, R. E., Hill, D. P., Lomax, J., Osumi-Sutherland, D., et al. (2014). TermGenie - a web-application for pattern-based ontology class generation. *J. Biomed. Semantics* 5, 48. doi: 10.1186/2041-1480-5-48

Dolin, R. H., Huff, S. M., Rocha, R. A., Spackman, K. A., and Campbell, K. E. (1998). Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. *J. Am. Med. Inform. Assoc.* 5 (2), 203–213. doi: 10.1136/jamia.1998.0050203

Dupuch, M., Dupuch, L., Hamon, T., and Grabar, N. (2014). Exploitation of semantic methods to cluster pharmacovigilance terms. *J. Biomed. Semantics* 5 (1), 18. doi: 10.1186/2041-1480-5-18

Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., et al., (2006). "Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. ," in *Mayo Clinic Proceedings*, vol. 81. (Elsevier), 741–748. doi: 10.4065/81.6.741

Emani, C. K., Da Silva, C. F., Fiès, B., Ghodous, P., and Bourdeau, M., (2015). "Automated Semantic Enrichment of Ontologies in the Construction Domain," in *Proc. of the 32nd CIB W78 Conference 2015, 27th-29th,* Eindhoven, Netherlands.

Fernández-Breis, J. T., Chiba, H., Legaz-García Mdel, C., and Uchiyama, I. (2016). The Orthology Ontology: development and applications. *J. Biomed. Semantics* 7 (1), 34. doi: 10.1186/s13326-016-0077-x

Fernández-López, M., Gómez-Pérez, A., and Juristo, N., (1997). Methontology: from ontological art towards ontological engineering. Proc. AAAI Spring Symp. Series, AAAI Press, Menlo Park, Calif., 33–40.

Fung, K. W., Hole, W. T., Nelson, S. J., Srinivasan, S., Powell, T., and Roth, L. (2005). Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. *J. Am. Med. Inform. Assoc.* 12 (4), 486–494. doi: 10.1197/jamia.M1767

Fung, K. W., Bodenreider, O., Aronson, A. R., Hole, W. T., and Srinivasan, S. (2007). Combining lexical and semantic methods of inter-terminology mapping using the UMLS. *Stud. Health Technol. Inform.* 129 (Pt 1), 605–609. doi: 10.3233/978-1-58603-774-1-605

Fung, K. W., Xu, J., Ameye, F., Gutiérrez, A. R., and D'Havé, A. (2017). Leveraging Lexical Matching and Ontological Alignment to Map SNOMED CT Surgical Procedures to ICD-10-PCS. *AMIA Annu. Symp. Proc.* 2016, 570–579. eCollection 2016.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In International semantic web conference pp. 262–276. Springer, Berlin, Heidelberg.

Gavankar, C., Kulkarni, A., Fang Li, Y., and Ramakrishnan. (2012). Enriching an academic knowledge base using linked open data In: Proceedings of Workshop on Speech and Language Processing Tools in Education in 24th International Conference on Computational Linguistics, 51–60.

Giannangelo, K., and Millar, J. (2012). Mapping SNOMED CT to ICD-10. *Stud. Health Technol. Inform.* 180, 83–87. doi: 10.3233/978-1-61499-101-4-83

Guo, A., Racz, R., Hur, J., Lin, Y., Xiang, Z., Zhao, L., et al. (2016). Ontology-based collection, representation and analysis of drug-associated neuropathy adverse events. *J. Biomed. Semantics* May 217, 29. doi: 10.1186/s13326-016-0069-x

Hansen, R. A., Gartlehner, G., Powell, G. E., and Sandler, R. S. (2007). Serious adverse events with infliximab: analysis of spontaneously reported adverse events. *Clin. Gastroenterol. Hepatol.* 5 (6), 729–735. doi: 10.1016/j.cgh.2007.02.016

Hartel, F. W., de Coronado, S., Dionne, R., Fragoso, G., and Golbeck, J. (2005). Modeling a description logic vocabulary for cancer research. *J. Biomed. Inform.* 38 (2), 114–129. doi: 10.1016/j.jbi.2004.09.001

Hauben, M., Patadia, V. K., and Goldsmith, D. (2006). What counts in data mining? *Drug Saf.* 29 (10), 827–832. doi: 10.2165/00002018-200629100-00001

He, Y., Sarntivijai, S., Lin, Y., Xiang, Z., Guo, A., Zhang, S., et al. (2014). OAE: the ontology of adverse events. *J. Biomed. Semantics* 2014, 5:29. doi: 10.1186/2041-1480-5-29

He, Y., Xiang, Z., Zheng, J., Lin, Y., Overton, J. A., and Ong, E. (2018). The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *J. Biomed. Semantics* 9 (1), 3. doi: 10.1186/s13326-017-0169-2

Henegar, C., Bousquet, C., Lillo-Le Louët, A., Degoulet, P., and Jaulent, M. C. (2006). Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. *Comput. Biol. Med.* 36 (7-8), 748–767. doi: 10.1016/j.compbiomed.2005.04.009

Huang, J., Dang, J., Borchert, G. M., Eilbeck, K., Zhang, H., Xiong, M., et al. (2014). OMIT: dynamic, semi-automated ontology development for the microRNA domain. *PLoS One* 9 (7), e100855. doi: 10.1371/journal.pone.0100855

Iavindrasana, J., Bousquet, C., and Jaulent, M. C. (2006). Knowledge acquisition for computation of semantic distance between WHO-ART terms. *Stud. Health Technol. Inform.* 124, 839–844. doi: 10.3233/978-1-58603-647-8-839

ICH Working Group. (2018). MedDRA® data retrieval and presentation: points to consider. Release 4.15 Based on MedDRA Version 21.0 MedDRA® TERM SELECTION: POINTS TO CONSIDER. ICH-Endorsed Guide for MedDRA Users. https://www.meddra.org/sites/default/files/guidance/file/000157_datretptc_r3_15_mar2018.pdf.

Ingenerf, J., and Giere, W. (1998). Concept-oriented standardization and statistics-oriented classification: continuing the classification versus nomenclature controversy. *Methods Inf. Med.* 37 (4-5), 527–539. doi: 10.1055/s-0038-1634544

Jimeno-Yepes, A., Jiménez-Ruiz, E., Berlanga-Llavori, R., and Rebholz-Schuhmann, D. (2009). Reuse of terminological resources for efficient ontological engineering in Life Sciences. *BMC Bioinformatics* 10 (10), S4. doi: 10.1186/1471-2105-10-S10-S4

Judkins, J., Tay-Sontheimer, J., Boyce, R. D., and Brochhausen, M. (2018). Extending the DIDEO ontology to include entities from the natural product drug interaction domain of discourse. *J. Biomed. Semantics* 9 (1), 15. doi: 10.1186/s13326-018-0183-z

Khorrami, F., Ahmadi, M., and Sheikhtaheri, A. (2018). Evaluation of SNOMED CT Content Coverage: a systematic literature review. *Stud. Health Technol. Inform.* 248, 212–219. doi: 10.3233/978-1-61499-858-7-212

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods Inf. Med.* 32 (4), 281–291. doi: 10.1055/s-0038-1634945

Liu, K., Hogan, W. R., and Crowley, R. S. (2011). Natural Language Processing methods and systems for biomedical ontology learning. *J. Biomed. Inform.* 44 (1), 163–179. doi: 10.1016/j.jbi.2010.07.006

Maedche, A., and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intell. Syst.* 16 (2), 72–79. doi: 10.1109/5254.920602

Marcos, E., Zhao, B., and He, Y. (2013). The Ontology of Vaccine Adverse Events (OVAE) and its usage in representing and analyzing adverse events associated with US-licensed human vaccines. *J. Biomed. Semantics* 4, 40. doi: 10.1186/2041-1480-4-40

Mazo, C., Salazar, L., Corcho, O., Trujillo, M., and Alegre, E. (2017). A histological ontology of the human cardiovascular system. *J. Biomed. Semantics* 8 (1), 47. doi: 10.1186/s13326-017-0158-5

McKnight, L. K., Elkin, P. L., Ogren, P. V., and Chute, C. G. (1999). Barriers to the clinical implementation of compositionality. *Proc. AMIA Symp.*, 320–324.

Merabti, T., Letord, C., Abdoune, H., Lecroq, T., Joubert, M., and Darmoni, S. J. (2009). Projection and inheritance of SNOMED CT relations between MeSH terms. *Stud. Health Technol. Inform.* 150, 233–237. doi: 10.3233/978-1-60750-044-5-233

Mozzicato, P. (2007). StandardisedMedDRA Queries. *Drug Saf.* 30 (7), 617–619. doi: 10.2165/00002018-200730070-00009

Nadkarni, P. M., and Darer, J. A. (2010). Determining correspondences between high frequency MedDRA concepts and SNOMED: a case study. *BMC Med. Inform. Deci. Mak.* 10. doi: 10.1186/1472-6947-10-66

Nikiema, J. N., Jouhet, V., and Mougin, F. (2017). Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *J. Biomed. Inform.* 74, 46–58. doi: 10.1016/j.jbi.2017.08.013

Nyström, M., Vikström, A., Nilsson, G. H., Ahlfeldt, H., and Orman, H. (2010). Enriching a primary health care version of ICD-10 using SNOMED CT mapping. *J. Biomed. Semantics* 1 (1), 7. doi: 10.1186/2041-1480-1-7

Oliveira, D., and Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *J. Biomed. Semantics* 9 (1), 1. doi: 10.1186/s13326-017-0171-8

Park, H., and Hardiker, N. (2009). Clinical terminologies: a solution for semantic interoperability. *J. Korean Soc. Med. Inform.* 15 (1), 1–11. doi: 10.4258/jksmi.2009.15.1.1

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E., (2011). "Ontology population and enrichment: State of the art," in *Knowledge-driven multimedia information extraction and ontology evolution*, Springer-Verlag, 134–166. doi: 10.1007/978-3-642-20795-2_6

Petrova, A., Ma, Y., Tsatsaronis, G., Kissa, M., Distel, F., Baader, F., et al. (2015). Formalizing biomedical concepts from textual definitions. *J. Biomed. Semantics* 6, 22. doi: 10.1186/s13326-015-0015-3

Presutti, V., and Gangemi, A. (2008). Content ontology design patterns as practical building blocks for web ontologies. In International Conference on Conceptual Modeling pp. 128–141. Springer, Berlin, Heidelberg.

Quesada-Martínez, M., Mikroyannidi, E., Fernández-Breis, J. T., and Stevens, R. (2015). Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. *Artif Intell Med.* Sep65 (1), 35–48. doi: 10.1016/j.artmed.2014.09.003

Roldán-García, M. D., García-Godoy, M. J., and Aldana-Montes, J. F. (2016). Dione: an OWL representation of ICD-10-CM for classifying patients' diseases. *J. Biomed. Semantics* 7 (1), 62. doi: 10.1186/s13326-016-0105-x

Qawasmeh, O., Lefrançois, M., Zimmermann, A., and Maret, P., (2018). Improved categorization of computer-assisted ontology construction systems: focus on bootstrapping capabilities. In Extended semantic web conference (ESWC2018). doi: 10.1007/978-3-319-98192-5_12

Rector, A., and Brandt, S. (2008). Why Do It the Hard Way? The Case for an Expressive Description Logic for SNOMED. *J. Am. Med. Inform. Assoc.* 15 (6), 744–751. doi: 10.1197/jamia.M2797

Rector, A., and Iannone, L. (2012). Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *J. Biomed. Inform.* 45 (2), 199–209. doi: 10.1016/j.jbi.2011.10.002

Riga, M., Mitzias, P., Kontopoulos, E., and Kompatsiaris, I., (2017). PROPheT–Ontology Population and Semantic Enrichment from Linked Data Sources. In International Conference on Data Analytics and Management in Data Intensive Domains, Springer, Cham 157–168. doi: 10.1007/978-3-319-96553-6_12

Rogers. (2005). Using Medical Terminologies. http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology-using.htm.

Rossi Mori, A., Consorti, F., and Galeazzi, E. (1998). Standards to support development of terminological systems for healthcare telematics. *Methods Inf. Med.* 37 (4-5), 551–563. doi: 10.1055/s-0038-1634542

Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L., and Brown, S. H. (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J. Am. Med. Inform. Assoc.* 13 (3), 277–288. doi: 10.1197/jamia.M1957

Sánchez, D., and Moreno, A. (2008). Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.* 64 (3), 600–623. doi: 10.1016/j.datak.2007.10.001

Sarntivijai, S., Xiang, Z., Shedden, K. A., Markel, H., Omenn, G. S., Athey, B. D., et al., (2012). Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines. doi: 10.1371/journal.pone.0049941

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., et al. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40 (D1), D940–D946. doi: 10.1093/nar/gkr972

Schulz, S., and Jansen, L. (2013). Formal ontologies in biomedical knowledge representation. *Yearb Med. Inform.* 8 (1), 132–146. doi: 10.1055/s-0038-1638845

Schulz, S., and Hahn, U. (2001). Medical knowledge reengineering–converting major portions of the UMLS into a terminological knowledge base. *Int. J. Med. Inform.* 64(2–3):207-21. doi: 10.1016/S1386-5056(01)00201-5

Schulz, S., Rodrigues, J. M., Rector, A., and Chute, C. G. (2017). Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Stud. Health Technol. Inform.* 245,940–944. doi 10.3233/978-1-61499-830-3-940

Serra, I., Girardi, R., and Novais, P. (2014). Evaluating techniques for learning non-taxonomic relationships of ontologies from text. *Expert Syst. Appl.* 41 (11), 5201–5211. doi: 10.1016/j.eswa.2014.02.042

Sheth, A., Ramakrishnan, C., and Thomas, C. (2005). Semantics for the semantic web: The implicit, the formal and the powerful. *Int. J. Semant Web Inf. Syst.* 1 (1), 1–18. doi: 10.4018/jswis.2005010101

Simperl, E. P. B., Tempich, C., and Sure, Y., (2006). "Ontocom: a cost estimation model for ontology engineering," in *International Semantic Web Conference* (Berlin, Heidelberg: Springer), 625–639. doi: 10.1007/11926078_45

Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W. L., and Wright, L. W. (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* 40 (1), 30–43. doi: 10.1016/j.jbi.2006.02.013

Souvignet, J., Declerck, G., Trombert, B., Rodrigues, J. M., Jaulent, M. C., and Bousquet, C. (2012). Evaluation of automated term groupings for detecting anaphylactic shock signals for drugs. *AMIA Annu. Symp. Proc.* 2012, 882–890.

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

Souvignet, J., Asfari, H., Lardon, J., Del Tedesco, E., Declerck, G., and Bousquet, C. (2016a). MedDRA automated term groupings using OntoADR: evaluation with upper gastrointestinal bleedings. *Expert Opin. Drug Saf.* 15 (9), 1153–1161. doi: 10.1080/14740338.2016.1206075

Souvignet, J., Declerck, G., Asfari, H., Jaulent, M. C., and Bousquet, C. (2016b). OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. *J. Biomed. Inform.* 63, 100–107. doi: 10.1016/j.jbi.2016.06.010

Souvignet, J., Declerck, G., Trombert-Paviot, B., Asfari, H., Jaulent, M. C., and Bousquet, C. (2019). Semantic Queries Expedite MedDRA Terms Selection Thanks to a Dedicated User Interface: A Pilot Study on Five Medical Conditions. *Front. Pharmacol.* 10, 50. doi: 10.3389/fphar.2019.00050

Spackman, K. A., Campbell, K. E., and Côté, R. A. (1997). SNOMED RT: a reference terminology for health care. *Proc AMIA Annu Fall Symp.* 640–4.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M., (2012). "The NeOn methodology for ontology engineering," in *Ontology engineering in a networked world* (Berlin, Heidelberg: Springer), 9–34. doi: 10.1007/978-3-642-24794-1_2

Sure, Y., Staab, S., and Studer, R., (2004). "On-to-knowledge methodology (OTKM)," in *Handbook on ontologies* (Berlin, Heidelberg: Springer), 117–132. doi: 10.1007/978-3-540-24750-0_6

Third, A. (2012). "Hidden semantics: what can we learn from the names in an ontology?," in *Proceedings of the Seventh International Natural Language Generation Conference, Association for Computational Linguistics*, 67–75.

Tiddi, I., Mustapha, N. B., Vanrompay, Y., and Aufaure, M. A., (2012). "Ontology learning from open linked data and web snippets," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (Berlin, Heidelberg: Springer), 434–443. doi: 10.1007/978-3-642-33618-8_59

Trifirò, G., Pariente, A., Coloma, P. M., Kors, J. A., Polimeni, G., Miremont-Salamé, G., et al. (2009). EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* 18 (12), 1176–1184. doi: 10.1002/pds.1836

van Damme, P., Quesada-Martínez, M., Cornet, R., and Fernández-Breis, J. T. (2018). From lexical regularities to axiomatic patterns for the quality assurance of biomedical terminologies and ontologies. *J. Biomed. Inform.* 84, 59–74. doi: 10.1016/j.jbi.2018.06.008

Vikström, A., Skånér, Y., Strender, L. E., and Nilsson, G. H. (2007). Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Med. Inform. Decis. Mak.* 2 (7), 9. doi: 10.1186/1472-6947-7-9

Villaverde, J., Persson, A., Godoy, D., and Amandi, A. (2009). Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. *Expert Syst. Appl.* 36 (7), 10288–10294. doi: 10.1016/j.eswa.2009.01.048

Wächter, T., and Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit. *Bioinformatics* 26 (12), i88–i96. doi: 10.1093/bioinformatics/btq188

Xiang, Z., Zheng, J., Lin, Y., and He, Y. (2015). Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. *J. Biomed. Semantics* 6, 4. doi: 10.1186/2041-1480-6-4